

Il Cervello come Ipotesi

Inferenza Bayesiana, Percezione e il Problema della Soggettività

Introduzione: la percezione come costruzione

Esiste una tentazione ricorrente nel pensiero ordinario, quella di concepire la percezione come una sorta di specchio fedele della realtà esterna: gli occhi raccolgono la luce, le orecchie catturano le vibrazioni dell'aria, e il cervello si limita a registrare passivamente ciò che il mondo offre ai sensi. Quest'immagine, seducente nella sua semplicità, è radicalmente falsa. O meglio: è vera soltanto in una misura talmente parziale da risultare fuorviante. La percezione non è ricezione ma costruzione; non è trascrizione ma inferenza; non è apertura immediata sul mondo ma elaborazione mediata da aspettative, modelli interni e strutture probabilistiche che il sistema nervoso ha raffinato nel corso dell'intera vita dell'individuo, e prima ancora nel corso dell'evoluzione della specie.

Questa monografia intende sviluppare e difendere una tesi precisa: il cervello opera fondamentalmente come un meccanismo inferenziale che genera ipotesi sulle cause nascoste delle sue stesse afferenze sensoriali, e lo fa attraverso un processo formalmente equivalente all'inferenza bayesiana. La tesi non è nuova nelle sue linee generali — essa ha radici nell'epistemologia kantiana, nell'opera di Hermann von Helmholtz sulla percezione inconscia, e nel lavoro più recente di Karl Friston sulla cosiddetta energia libera variazionale. Ciò che si propone qui è tuttavia qualcosa di distinto da una mera esposizione del paradigma esistente: si tratta di un'analisi critica che mette in luce le implicazioni filosofiche profonde di questa architettura computazionale, ne identifica i punti di tensione con la fenomenologia della coscienza, e propone una lettura originale del rapporto tra la struttura matematica dell'inferenza e il problema, ancora irrisolto, dell'esperienza soggettiva.

Il percorso argomentativo si articola in quattro movimenti. Nel primo si pongono le fondamenta matematiche: il teorema di Bayes, la nozione di divergenza di Kullback-Leibler, e il principio di energia libera variazionale vengono derivati e discussi con il rigore che la loro applicazione biologica richiede. Nel secondo si esamina come questo apparato formale si traduca in un'architettura neurale concreta, quella del cosiddetto predictive coding, e si analizzano le sue conseguenze per la filosofia della percezione. Nel terzo si affronta il problema più arduo: come questa macchina inferenziale possa dar luogo a qualcosa come la soggettività, l'esperienza in

prima persona, il carattere fenomenico del percepire. Nel quarto, infine, si formulano le critiche più stringenti al paradigma, identificando le sue lacune esplicative e i circoli che ne minacciano la coerenza.

1. Il Formalismo dell'Inferenza e la sua Traduzione Neurale

1.1 Il Teorema di Bayes come struttura fondamentale della cognizione

Ogni atto percettivo implica una forma di ragionamento inverso. Quando il sistema visivo elabora un pattern di illuminazione sulla retina, il problema che deve risolvere non è di tipo diretto — non si tratta di calcolare quale pattern luminoso corrisponde a un certo oggetto — bensì inverso: date le eccitazioni retiniche, qual è l'oggetto più plausibile nel mondo esterno che potrebbe averle causate? Questo problema dell'inversione — risalire dagli effetti alle cause — è precisamente il dominio del teorema di Bayes.

Nella sua forma elementare, il teorema afferma che la probabilità di un'ipotesi H dato un corpo di evidenza E è proporzionale al prodotto della verosimiglianza dell'evidenza sotto quell'ipotesi per la probabilità a priori dell'ipotesi stessa:

$$P(H | E) = [P(E | H) \cdot P(H)] / P(E)$$

dove $P(H | E)$ è la probabilità posteriore dell'ipotesi, $P(E | H)$ è la verosimiglianza, $P(H)$ è la probabilità a priori, e $P(E)$ è la probabilità marginale dell'evidenza, ovvero la somma — o l'integrale, nel caso continuo — su tutte le ipotesi possibili:

$$P(E) = \sum_i P(E | H_i) \cdot P(H_i)$$

La traduzione di questa struttura in termini cognitivi è immediata nella sua formulazione, benché straordinariamente ricca nelle sue implicazioni. Si interpreti H come un'ipotesi sullo stato del mondo esterno — la presenza di un volto in una certa posizione, la traiettoria di un oggetto in movimento, la natura di uno stimolo acustico — e si interpreti E come l'insieme delle eccitazioni sensoriali disponibili al sistema nervoso. Allora $P(H | E)$ rappresenta la credenza aggiornata del sistema circa lo stato del mondo dopo aver ricevuto l'input sensoriale, $P(E | H)$ rappresenta la capacità del modello interno di predire correttamente gli input sensoriali a partire da quella ipotesi sullo stato del mondo, e $P(H)$ rappresenta l'aspettativa a priori su quanto probabile sia quello stato del mondo prima ancora di ricevere qualunque informazione sensoriale nella situazione attuale.

Questa interpretazione bayesiana della percezione ha una conseguenza che merita di essere sottolineata con forza, perché contraddice radicalmente l'immagine intuitiva con cui si è aperto questo saggio: il cervello non elabora la realtà a partire dall'input sensoriale grezzo, procedendo dal basso verso l'alto per accumulare informazioni fino a raggiungere una rappresentazione del mondo. Al contrario, esso mantiene in ogni momento un modello probabilistico del mondo, una distribuzione di probabilità su stati possibili, e utilizza l'input sensoriale soltanto per aggiornare questo modello secondo la regola bayesiana. Il flusso causale primario, per quanto riguarda la percezione conscia, procede dall'alto verso il basso: dal modello all'input, non viceversa.

1.2 Il problema computazionale: l'intrattabilità del calcolo esatto

La bellezza del formalismo bayesiano si scontra tuttavia con un ostacolo di natura computazionale che è impossibile trascurare. Il calcolo del termine $P(E)$ — la probabilità marginale dell'evidenza — richiede di sommare su tutti gli stati del mondo possibili, che in qualunque contesto realistico costituiscono uno spazio di dimensionalità astronomica. Se si considera che il cervello deve gestire simultaneamente ipotesi sulla forma, il colore, la posizione, la velocità, l'identità e innumerevoli altre proprietà degli oggetti percepiti, è evidente che il calcolo esatto dell'inferenza bayesiana è computazionalmente intrattabile: nessun sistema fisico di dimensioni biologicamente plausibili potrebbe eseguirlo in tempo reale.

Questo non è un problema tecnico secondario: è una sfida fondamentale alla teoria. Se la mente opera secondo la logica bayesiana ma il calcolo esatto è impossibile, allora il cervello deve implementare qualche forma di approssimazione. La domanda diventa: quale? E con quale grado di precisione? La risposta più potente e matematicamente elegante a queste domande è offerta dal framework della inferenza variazionale.

1.3 L'inferenza variazionale e la distribuzione approssimante

L'idea centrale dell'inferenza variazionale consiste nel sostituire il calcolo diretto della distribuzione posteriore $P(H | E)$ — intrattabile — con la ricerca della distribuzione approssimante $Q(H)$ più simile possibile alla vera posteriore, all'interno di una famiglia di distribuzioni parametriche che si prestano a calcoli efficienti. Il criterio di "somiglianza" è misurato attraverso la divergenza di Kullback-Leibler, una funzione che quantifica quanto due distribuzioni di probabilità differiscano l'una dall'altra.

Per due distribuzioni discrete Q e P definite sullo stesso spazio di ipotesi, la divergenza di KL di Q rispetto a P è definita come:

$$D_{KL}(Q \parallel P) = \sum_h Q(h) \cdot \ln[Q(h) / P(h)]$$

Nel caso continuo, la somma diventa un integrale:

$$D_{KL}(Q \parallel P) = \int Q(h) \cdot \ln[Q(h) / P(h)] dh$$

Questa quantità gode di alcune proprietà matematiche che la rendono particolarmente adatta allo scopo. In primo luogo, $D_{\text{KL}}(Q \parallel P) \geq 0$ per qualunque coppia di distribuzioni, con uguaglianza se e solo se $Q = P$ quasi ovunque — ciò segue immediatamente dalla disuguaglianza di Jensen applicata alla funzione convessa $-\ln$. In secondo luogo, essa non è simmetrica: $D_{\text{KL}}(Q \parallel P) \neq D_{\text{KL}}(P \parallel Q)$ in generale, il che implica che la scelta della direzione in cui si minimizza la divergenza ha conseguenze non banali per il comportamento del sistema approssimante.

L'inferenza variazionale si propone dunque di minimizzare $D_{\text{KL}}(Q \parallel P(H \mid E))$ rispetto ai parametri che determinano Q . Il problema è che questo calcolo richiederebbe ancora di conoscere $P(H \mid E)$, che è precisamente ciò che si vuole approssimare. Si ha bisogno di una riformulazione che eviti questa circolarità.

1.4 L'energia libera variazionale: derivazione e interpretazione

La soluzione alla circolarità appena descritta nasce da una manipolazione algebrica elegante. Si parte dall'identità che lega la log-verosimiglianza dei dati alla divergenza di KL tra la distribuzione approssimante e la vera posteriore. Per un vettore di osservazioni $\mathbf{o}$ e un insieme di variabili latenti $\mathbf{s}$ che rappresentano gli stati del mondo, si può scrivere:

$$\ln P(\mathbf{o}) = F[Q] + D_{\text{KL}}(Q(\mathbf{s}) \parallel P(\mathbf{s} \mid \mathbf{o}))$$

dove $F[Q]$ è una funzione della distribuzione approssimante Q chiamata energia libera variazionale con segno negativo, ovvero il cosiddetto ELBO (Evidence Lower Bound):

$$\begin{aligned} F[Q] &= E_Q[\ln P(\mathbf{o}, \mathbf{s})] - E_Q[\ln Q(\mathbf{s})] \\ &= E_Q[\ln P(\mathbf{o}, \mathbf{s})] + H[Q] \end{aligned}$$

con $H[Q] = -E_Q[\ln Q(\mathbf{s})]$ che indica l'entropia di Shannon della distribuzione Q . Poiché $D_{\text{KL}} \geq 0$, segue immediatamente che:

$$F[Q] \leq \ln P(\mathbf{o})$$

cioè $F[Q]$ è un limite inferiore della log-evidenza. Massimizzare $F[Q]$ rispetto ai parametri di Q equivale a minimizzare $D_{\text{KL}}(Q(\mathbf{s}) \parallel P(\mathbf{s} \mid \mathbf{o}))$, avvicinando la distribuzione approssimante alla vera posteriore, e al tempo stesso a massimizzare un limite inferiore della log-evidenza del modello — vale a dire, a trovare il modello del mondo che rende più probabili le osservazioni effettive.

Nella letteratura di Friston, l'energia libera variazionale viene più spesso definita con segno opposto — come la quantità da minimizzare — e riscritta in forme equivalenti che illuminano aspetti diversi del suo significato. Una decomposizione particolarmente rivelatrice è la seguente:

$$F = D_{\text{KL}}(Q(s) \parallel P(s)) - E_Q[\ln P(o \mid s)]$$

Il primo termine, $D_{\text{KL}}(Q(s) \parallel P(s))$, misura quanto la distribuzione posteriore approssimata si discosta dalla distribuzione a priori sugli stati del mondo — è un termine di "costo" per la complessità del modello aggiornato, che penalizza le revisioni eccessive delle credenze a priori. Il secondo termine, $-E_Q[\ln P(o \mid s)]$, misura l'accuratezza predittiva del modello — è un termine di "precisione" che premia i modelli capaci di spiegare le osservazioni effettive. La minimizzazione dell'energia libera è quindi la ricerca di un equilibrio ottimale tra accuratezza e complessità: il sistema cerca il modello più semplice che spieghi adeguatamente i dati, il che rispecchia in modo preciso il principio del rasoio di Occam nella sua formulazione probabilistica.

Una seconda decomposizione equivalente è altrettanto illuminante:

$$\begin{aligned} F &= E_Q[\ln Q(s) - \ln P(s, o)] \\ &= E_Q[\ln Q(s) - \ln P(o \mid s) - \ln P(s)] \end{aligned}$$

Riscrivendo $\ln P(o \mid s)$ come la predizione del modello e $\ln Q(s) - \ln P(s)$ come la divergenza delle credenze dalla loro configurazione a priori, si vede come F si possa interpretare come la somma degli errori di predizione ponderati per la precisione attribuita a ciascun livello della gerarchia. Questa terza lettura sarà decisiva quando si affronterà l'architettura neurale del predictive coding.

1.5 La necessità del modello generativo

Prima di procedere, è indispensabile chiarire un elemento che spesso rimane implicito nelle esposizioni del paradigma: tutta la struttura fin qui delineata presuppone l'esistenza di un modello generativo. Un modello generativo è una rappresentazione interna della struttura causale del mondo — una specifica delle distribuzioni $P(s)$ sugli stati del mondo e $P(o \mid s)$ sulla relazione tra stati del mondo e osservazioni sensoriali — che permette al sistema di simulare come il mondo produrrebbe, in linea di principio, gli input sensoriali effettivi.

L'esistenza di tale modello è una delle ipotesi più impegnative del framework. Essa implica che il cervello non soltanto registra correlazioni statistiche tra input sensoriali, ma rappresenta esplicitamente o implicitamente le strutture causali che generano quegli input. Questa è una posizione epistemologica robusta, che si colloca nel solco di una lunga tradizione che va da Kant — con la sua dottrina delle forme a priori dell'intuizione e delle categorie dell'intelletto — a Helmholtz — con la sua teoria dell'inferenza inconscia — fino ai costruttivisti contemporanei come Richard Gregory.

È importante precisare che "modello generativo" non implica necessariamente rappresentazione esplicita e localizzabile in qualche struttura cerebrale identificabile. Il modello generativo potrebbe essere distribuito attraverso l'intera architettura della rete neurale, codificato nei pesi sinaptici e nelle connettività piuttosto che in specifiche

unità. Ma la sua esistenza funzionale è un requisito imprescindibile della teoria: senza un modello del modo in cui il mondo causa le sensazioni, non è possibile calcolare la verosimiglianza $P(o | s)$ e l'intera struttura dell'inferenza bayesiana collassa.

Questo punto merita una riflessione filosofica che vada oltre la neurobiologia. Se il cervello possiede un modello generativo del mondo, allora in un senso non banale esso "comprende" il mondo — non nel senso in cui si comprende una proposizione matematica, ma nel senso in cui si possiede una rappresentazione strutturale delle dipendenze causali che governano l'ambiente. La cognizione, in questa prospettiva, non è una mera raccolta di risposte adattive a stimoli ambientali, ma una forma di modellizzazione del mondo che ha struttura e contenuto propri, relativamente indipendenti dai dati sensoriali specifici che li hanno generati.

1.6 Struttura gerarchica e la moltiplicazione dei livelli di astrazione

La trattazione svolta finora ha considerato per semplicità un solo livello di stati latenti s . Il cervello reale, tuttavia, opera attraverso una gerarchia di aree corticali di ordine crescente, e questa struttura anatomica trova una corrispondenza diretta nell'estensione gerarchica del modello generativo. A ogni livello della gerarchia corrispondono stati latenti di diverso grado di astrazione: ai livelli inferiori si trovano le caratteristiche locali e transitorie — orientamento dei bordi, frequenze spaziali, valori di lucentezza — mentre ai livelli superiori si trovano le strutture più stabili e astratte — categorie di oggetti, agenti, relazioni causali.

In un modello generativo gerarchico con L livelli, la distribuzione congiunta sugli stati del mondo si fattorizza come:

$$P(o, s^1, s^2, \dots, s^L) = P(o | s^1) \cdot P(s^1 | s^2) \cdot P(s^2 | s^3) \cdot \dots \cdot P(s^{L-1} | s^L) \cdot P(s^L)$$

dove ogni termine $P(s^{l-1} | s^l)$ descrive come gli stati al livello l generino le distribuzioni sugli stati al livello $l-1$, fino a che gli stati al livello più basso s^1 generino le osservazioni sensoriali o attraverso $P(o | s^1)$. Questo schema gerarchico permette di modellare strutture causali di complessità arbitraria attraverso la composizione di modelli locali relativamente semplici.

L'inferenza variazionale in questo contesto gerarchico porta alla fattorizzazione della distribuzione approssimante secondo la cosiddetta assunzione di campo medio:

$$Q(s^1, s^2, \dots, s^L) = Q(s^1) \cdot Q(s^2) \cdot \dots \cdot Q(s^L)$$

che rende il calcolo trattabile al prezzo di ignorare le correlazioni tra livelli diversi nella distribuzione posteriore approssimata. Questa approssimazione ha conseguenze importanti: in particolare, essa tende a produrre distribuzioni posteriori eccessivamente concentrate rispetto alle vere posteriori, un fenomeno noto come "variational collapse" che potrebbe avere rilevanza per comprendere certe forme di elaborazione rigida dell'informazione osservate in alcune condizioni cliniche.

1.7 Il ruolo della precisione: ponderare le fonti di informazione

Un aspetto del framework che assume importanza crescente nelle applicazioni sia neuroscientifiche sia cliniche è la nozione di precisione — ovvero la precisione statistica, reciproco della varianza, attribuita agli errori di predizione a ciascun livello della gerarchia. In un modello gaussiano, la distribuzione degli errori di predizione e a un certo livello è caratterizzata da una media nulla e da una matrice di covarianza Σ :

$$\epsilon \sim N(0, \Sigma)$$

La precisione associata a questo livello è la matrice inversa $\Pi = \Sigma^{-1}$. Quando il sistema minimizza l'energia libera variazionale in presenza di errori di predizione ponderati per la precisione, il contributo di ciascun livello alla minimizzazione complessiva è moltiplicato per la sua precisione:

$$F \approx \sum_l (1/2) \cdot \epsilon^l{}^T \cdot \Pi^l \cdot \epsilon^l + \text{costanti}$$

dove l'indice l percorre i livelli della gerarchia. Questo meccanismo di ponderazione per la precisione è computazionalmente equivalente all'attenzione selettiva: aumentare la precisione attribuita a un certo livello significa dare maggiore peso agli errori di predizione provenienti da quel livello, il che si traduce in una maggiore sensibilità alle discrepanze tra predizioni e segnali sensoriali in quella regione dello spazio delle caratteristiche. L'attenzione, in questa prospettiva, non è un meccanismo distinto dalla percezione ma è intrinsecamente incorporata nella struttura stessa dell'inferenza: è la modulazione della precisione.

Questa identificazione dell'attenzione con la precisione ha implicazioni terapeutiche e cliniche notevoli. La psicosi, per esempio, potrebbe essere reinterpretata non come un difetto nella costruzione di rappresentazioni errate del mondo, ma come una perturbazione del meccanismo di ponderazione della precisione: se la precisione attribuita ai segnali sensoriali diventa abnormemente elevata rispetto alla precisione delle aspettative a priori, il sistema sarà incapace di "smorzare" gli errori di predizione nei modi abituali, e attribuirà salienza aberrante a coincidenze e dettagli irrilevanti, generando credenze bizzarre nel tentativo di spiegarli. Similmente, certe forme di ansia potrebbero essere caratterizzate da una precisione eccessivamente elevata delle aspettative a priori di minaccia, che porta il sistema a "vedere" pericoli laddove non esistono, non perché le inferenze siano malformate ma perché sono eseguite a partire da priori distorti.

2. 1 Predictive Coding e la Filosofia della Percezione

2.1 Dall'inferenza variazionale all'architettura neurale

Il passaggio dal formalismo matematico all'architettura biologica è uno dei momenti più fecondi e, al tempo stesso, più rischiosi dell'intero paradigma. Fecondo perché produce previsioni verificabili sull'organizzazione del sistema nervoso; rischioso perché la corrispondenza tra strutture matematiche astratte e substrati neurali concreti non è mai biunivoca, e l'entusiasmo per le analogie può indurre a sopravvalutare la conferma empirica.

L'architettura del predictive coding — o codifica predittiva — è la proposta più articolata di implementazione neurale del framework variazionale. Nella sua versione standard, essa attribuisce a popolazioni distinte di neuroni corticali la funzione di rappresentare rispettivamente le predizioni discendenti del modello generativo e gli errori di predizione ascendenti. Ripercorriamo la derivazione di questa architettura partendo dalla condizione di ottimalità che si ottiene quando si minimizza l'energia libera rispetto alle attivazioni neurali.

Supponiamo che la distribuzione approssimante $Q(s^l)$ a ciascun livello l della gerarchia sia gaussiana con media μ^l e precisione Π^l :

$$Q(s^l) = N(\mu^l, (\Pi^l)^{-1})$$

Il modello generativo impone che gli stati al livello l siano generati dagli stati al livello superiore $l+1$ attraverso una funzione non lineare g :

$$s^l = g(s^{l+1}) + w, \quad w \sim N(0, (\Pi^l)^{-1})$$

dove w rappresenta il rumore di processo con precisione Π^l . In questo caso, l'energia libera variazionale può essere riscritta come:

$$F = \sum_l (1/2) \cdot \xi^l \cdot \Pi^l \cdot \xi^l - (1/2) \ln \det(\Pi^l) + \text{costante}$$

dove $\xi^l = \mu^l - g(\mu^{l+1})$ è l'errore di predizione al livello l , ovvero la differenza tra la stima corrente degli stati a quel livello e la predizione generata dagli stati al livello superiore. La condizione di ottimalità — che si ottiene annullando il gradiente di F rispetto a μ^l — dà luogo a un'equazione differenziale che descrive la dinamica temporale delle rappresentazioni:

$$\dot{\mu}^l = -\partial F / \partial \mu^l = \Pi^l \cdot \xi^l - (\partial g(\mu^l) / \partial \mu^l)^T \cdot \Pi^{l+1} \cdot \xi^{l+1}$$

La lettura di questa equazione rivela una struttura elegante: la variazione temporale delle rappresentazioni al livello l è determinata da due contributi di segno opposto. Il

primo, $\Pi^l \cdot \xi^l$, spinge μ^l nella direzione che riduce l'errore di predizione dal livello superiore — è il contributo del segnale discendente. Il secondo, che coinvolge il Jacobiano $\partial g / \partial \mu^l$ e l'errore al livello inferiore ξ^{l-1} , spinge μ^l nella direzione che riduce l'errore di predizione verso il livello inferiore — è il contributo del segnale ascendente. Il sistema raggiunge l'equilibrio quando i due contributi si bilanciano, cioè quando le rappresentazioni sono al tempo stesso consistenti con le aspettative provenienti dall'alto e capaci di spiegare i dati provenienti dal basso.

2.2 Neuroni di predizione ed errore: una distinzione funzionale

Dalla dinamica appena derivata emergono due classi funzionali di unità corticali. Le unità che rappresentano le medie μ^l — le stime correnti degli stati del mondo — sono convenzionalmente chiamate neuroni di rappresentazione o neuroni di stato. Le unità che rappresentano gli errori di predizione pesati per la precisione, $\xi^l \cdot \Pi^l$ (spesso indicati come ϵ^l), sono i neuroni di errore. La comunicazione ascendente nel sistema corticale è prevalentemente portata dai neuroni di errore — che segnalano ai livelli superiori le discrepanze non spiegate dal modello — mentre le connessioni discendenti trasmettono le predizioni dei neuroni di rappresentazione ai livelli inferiori, dove vengono confrontate con le rappresentazioni locali per generare nuovi errori.

Questa architettura trova una corrispondenza anatomica parziale nell'organizzazione laminare della corteccia cerebrale. Le cellule piramidali degli strati superficiali della corteccia — in particolare lo strato II e III — sono i principali candidati per il ruolo di neuroni di errore che proiettano verso aree di ordine superiore attraverso connessioni feedforward. Le cellule degli strati profondi — in particolare lo strato V e VI — proiettano invece verso aree di ordine inferiore attraverso connessioni di feedback, e sono candidate naturali per il ruolo di neuroni di rappresentazione che trasmettono le predizioni discendenti. Questa identificazione, proposta originariamente da Mumford e sviluppata da Friston, è suggestiva ma rimane soggetta a revisione alla luce dei dati neurofisiologici, che mostrano un'organizzazione più complessa e sovrapposta di quanto la dicotomia semplice suggerirebbe.

2.3 L'inversione del problema: la percezione come soppressione dell'errore

Una delle conseguenze più controintuitive e filosoficamente più fertili del framework è la seguente: una percezione di successo non è quella che genera la risposta neurale più intensa, bensì quella che genera la risposta più silenziosa. Quando il modello generativo predice correttamente l'input sensoriale, gli errori di predizione ai livelli bassi della gerarchia tendono a zero, e l'attività dei neuroni di errore si riduce. Il percepito — la rappresentazione cosciente dell'oggetto percepito — corrisponde all'insieme di stati del modello generativo che hanno ridotto al minimo gli errori di predizione residui.

Questa inversione è di grande importanza per reinterpretare un ampio corpus di dati sperimentali. L'adattamento neurale, per esempio — il fenomeno per cui i neuroni

riducono la loro risposta a uno stimolo ripetuto — si spiegherebbe naturalmente come il progressivo affinamento delle predizioni: uno stimolo ripetuto diventa sempre più anticipato dal modello, gli errori di predizione si riducono, e i neuroni di errore si silenzieranno di conseguenza. Le illusioni visive, d'altra parte, corrisponderebbero a situazioni in cui il modello generativo è talmente ben consolidato — le aspettative a priori sono talmente precise — da sopraffare l'evidenza sensoriale contraddittoria: il sistema "vede" ciò che si aspetta di vedere piuttosto che ciò che è fisicamente presente, perché il costo dell'errore di predizione a livello sensoriale è inferiore al costo di aggiornare una rappresentazione ad alta precisione.

Il sogno, in questa cornice, diventa un caso estremo di dominanza del modello generativo: in assenza di input sensoriale affidabile — la precisione dei segnali sensoriali scende drasticamente durante il sonno REM — il sistema opera quasi esclusivamente sulla base del suo modello interno, generando esperienze che riflettono la struttura delle aspettative apprese piuttosto che la struttura del mondo esterno. Non è privo di significato che i sogni abbiano spesso una coerenza narrativa parziale, pur essendo fenomenologicamente indistinguibili dalla veglia per colui che li vive: la coerenza riflette la struttura del modello generativo, l'assenza di critica riflessiva riflette l'assenza del segnale di errore che nella veglia costituisce il principale meccanismo di correzione delle credenze.

2.4 Helmholtz, Kant, e la preistoria del cervello predittivo

Prima di procedere verso le implicazioni filosofiche più impegnative, vale la pena situare il paradigma nel suo contesto intellettuale, perché la genealogia delle idee è spesso rivelatrice di tensioni teoriche che rischiano altrimenti di passare inosservate.

Hermann von Helmholtz, nella sua monumentale opera sulla fisiologia ottica del 1867, propose che la percezione fosse un processo di inferenza inconscia: gli organi di senso forniscono dati ambigui che devono essere interpretati alla luce di ipotesi sul mondo esterno, e questa interpretazione avviene automaticamente e al di fuori della consapevolezza del percipiente. La visione di profondità, per esempio, non è direttamente disponibile nell'immagine bidimensionale che cade sulla retina — essa deve essere ricostruita a partire da indizi come la disparità binoculare, la prospettiva lineare e il movimento relativo, secondo regole che il sistema visivo apprende dall'esperienza. Helmholtz parlava esplicitamente di "ipotesi" sulla natura degli oggetti percepiti, e usava il linguaggio della probabilità — sia pur in modo intuitivo piuttosto che formale — per descrivere come il sistema visivo selezionasse le ipotesi più plausibili.

Il legame con Kant è più complesso e più profondo. La dottrina kantiana delle forme a priori dell'intuizione — lo spazio e il tempo come strutture imposte dalla mente agli oggetti dell'esperienza, piuttosto che proprietà degli oggetti in sé — trova un'eco nel fatto che il modello generativo del cervello predittivo non è una tabula rasa ma una struttura con vincoli a priori che riflettono la storia evolutiva e ontogenetica dell'organismo. Le aspettative a priori $P(s)$ non sono arbitrarie ma riflettono le regolarità statistiche del mondo in cui l'organismo vive — un mondo tridimensionale

in cui gli oggetti si muovono in modo continuo, in cui l'illuminazione viene generalmente dall'alto, in cui i volti hanno una struttura caratteristica. Queste aspettative costituiscono una sorta di "a priori empirico" — una struttura che precede l'esperienza attuale ma è stata formata attraverso l'esperienza passata, sia filogenetica sia ontogenetica.

Tuttavia, c'è una differenza cruciale tra la posizione kantiana e quella del cervello predittivo che non deve essere occultata. Per Kant, le forme a priori dell'intuizione e le categorie dell'intelletto sono necessarie e universali — esse costituiscono le condizioni di possibilità di qualunque esperienza, e non potrebbero essere altrimenti. Per il framework bayesiano, le aspettative a priori sono contingenti e modificabili: esse riflettono le statistiche dell'ambiente in cui l'organismo si è sviluppato, e potrebbero in linea di principio essere diverse se l'ambiente fosse diverso. Il "a priori" del cervello predittivo è un a priori biologico e statistico, non trascendentale. Questa differenza ha implicazioni profonde per il problema della oggettività della conoscenza, su cui si tornerà nella terza parte.

2.5 Il contenuto della percezione: internismo o esternismo?

L'architettura del predictive coding solleva con urgenza rinnovata uno dei dibattiti centrali della filosofia contemporanea della mente: il contrasto tra internismo ed esternismo riguardo al contenuto dei stati mentali.

L'internista sostiene che il contenuto di uno stato percettivo è interamente determinato dalla struttura interna del sistema percipiente — dalle sue rappresentazioni, dai suoi stati neurali, dai suoi modelli. Due sistemi internamente identici avrebbero necessariamente gli stessi stati mentali, indipendentemente dall'ambiente in cui si trovano. L'esternista, al contrario, sostiene che il contenuto è parzialmente determinato dalla relazione causale e storica tra il sistema e il mondo esterno — da Putnam con i suoi "gemelli della Terra" a Tyler Burge con i suoi esperimenti mentali sull'artrite, la tradizione esternista ha accumulato argomentazioni potenti a favore dell'idea che "i significati non stanno nella testa".

Il framework del cervello predittivo sembra a prima vista collocarsi nel campo internista: ciò che determina la percezione è il modello generativo interno, le aspettative a priori, la struttura delle distribuzioni di probabilità mantenute dal sistema. Il segnale sensoriale gioca un ruolo, ma è un ruolo di aggiornamento — di perturbazione di un sistema che ha già una struttura ricca e autonoma. Due sistemi con modelli generativi identici avrebbero la stessa esperienza percettiva di fronte agli stessi input sensoriali, indipendentemente dalla storia causale che ha prodotto quei modelli.

Tuttavia, questa conclusione è prematura. Il modello generativo stesso non cade dal cielo: è il prodotto di un processo di apprendimento che si è svolto in relazione causale con un ambiente specifico. Le distribuzioni a priori $P(s)$ codificano le statistiche di quell'ambiente. In un senso diacronico, quindi, il contenuto del modello è esternamente determinato — anche se, in un senso sincronico, è interno al sistema. Si

tratta di una forma di "esternismo storico" compatibile con un "internismo sincronico": il contenuto delle rappresentazioni è determinato dalla storia causale del sistema con il suo ambiente, ma una volta fissato quel contenuto, la percezione è governata dalla struttura interna.

Questa posizione ha affinità con quella che Andy Clark e David Chalmers hanno chiamato "mente estesa": se il confine tra sistema cognitivo e ambiente non è fisso, allora la domanda sul "dove" si trova il contenuto perde parte della sua presa. Il framework predittivo suggerisce che il sistema cognitivo non è il cervello in isolamento ma il ciclo dinamico di percezione e azione attraverso cui il sistema costruisce e verifica continuamente il suo modello del mondo — un circolo ermeneutico biologicamente istanziato.

2.6 La percezione come inferenza alla migliore spiegazione

Un modo alternativo di caratterizzare l'epistemologia implicita nel framework è attraverso il concetto di inferenza alla migliore spiegazione, noto anche come abduzione nel senso di Peirce. L'abduzione è quella forma di ragionamento per cui, dato un insieme di fatti osservati, si conclude che la migliore spiegazione di quei fatti è vera — o almeno, che è ragionevole credere che lo sia.

Nel contesto del cervello predittivo, ogni atto percettivo è abduttivo: il sistema seleziona tra le ipotesi sullo stato del mondo quella che meglio spiega — nel senso tecnico di rendere più probabili — gli input sensoriali effettivamente ricevuti. La struttura bayesiana fornisce un criterio preciso per questa selezione: la migliore spiegazione è quella che massimizza la probabilità posteriore $P(H \mid E)$, bilanciando la capacità esplicativa $P(E \mid H)$ con la plausibilità a priori $P(H)$.

Questa lettura abduttiva della percezione ha conseguenze interessanti per la comprensione degli errori percettivi. Le illusioni non sono fallimenti del sistema percettivo ma risposte ottimali a situazioni in cui i segnali sensoriali sono ambigui — la migliore inferenza dato l'input disponibile, alla luce delle aspettative a priori, porta a una rappresentazione che diverge dalla realtà fisica. La figura di Müller-Lyer, con le sue linee di lunghezza uguale che sembrano diverse, è la risposta ottimale di un sistema che ha appreso le regolarità prospettiche del mondo tridimensionale a stimoli bidimensionali costruiti ad arte per violare quelle regolarità. Il sistema non sbaglia — applica le regole giuste nel contesto sbagliato, il che è esattamente ciò che farebbe un agente razionale bayesiano con le aspettative a priori appropriate.

2.7 Azione come inferenza: il principio di energia libera nella sua generalità

Finora si è considerato il cervello come un sistema che minimizza l'energia libera variazionale aggiornando le sue rappresentazioni interne — modificando μ^l in modo da ridurre gli errori di predizione. Ma il framework di Friston propone un'estensione radicale: la stessa struttura variazionale governa non soltanto la percezione ma anche l'azione.

L'idea, in apparenza semplice, è ricca di implicazioni: invece di aggiornare le rappresentazioni interne per farle corrispondere all'input sensoriale, il sistema può ridurre gli errori di predizione modificando l'input sensoriale stesso — cioè agendo sul mondo per rendere il mondo conforme alle proprie predizioni. Se il modello generativo predice che il braccio si troverà in una certa posizione, e la proprioccezione segnala che il braccio è in una posizione diversa, si può ridurre l'errore di predizione propriocettivo sia aggiornando la predizione sia muovendo il braccio fino alla posizione predetta.

In questa prospettiva, i riflessi spinali e i movimenti volontari complessi sono istanze dello stesso meccanismo fondamentale: la soppressione degli errori di predizione sensoriale attraverso l'azione. I movimenti volontari si distinguono dai riflessi non per il meccanismo ma per il livello della gerarchia a cui le predizioni vengono generate: nei movimenti volontari, le predizioni vengono generate da livelli superiori della gerarchia e trasmesse verso il basso fino al livello propriocettivo, dove vengono implementate dall'apparato motorio. I riflessi sono circuiti a basso livello che minimizzano rapidamente gli errori propriocettivi locali senza coinvolgere i livelli superiori.

Questa unificazione di percezione e azione sotto un unico principio variazionale — comunemente indicata come principio di energia libera nella sua generalità, o "active inference" — è uno dei contributi teorici più originali del paradigma. Essa suggerisce che la distinzione tradizionale tra sistemi sensoriali e sistemi motori è, a un livello di analisi sufficientemente astratto, una distinzione di grado piuttosto che di natura: entrambi implementano la minimizzazione di errori di predizione, rispettivamente attraverso l'aggiornamento delle credenze e attraverso la modificazione del segnale sensoriale in ingresso.

2.8 Il problema dell'intentionalità originaria

Il quadro delineato nelle sezioni precedenti potrebbe sembrare risolvere uno dei problemi centrali della filosofia della mente — quello dell'intentionalità, ovvero del "essere di" o "riferirsi a" qualcosa di esterno da parte degli stati mentali. Le rappresentazioni interne del modello generativo sembrano "riguardare" il mondo esterno nel senso più preciso: esse sono costruite per predire gli stati del mondo, e il loro contenuto è determinato dalla struttura causale delle relazioni tra il sistema e il suo ambiente.

Ma questo sarebbe un errore. Il problema dell'intentionalità originaria — sollevato in modo acuto da John Searle — non è soltanto la questione di come le strutture interne siano correlate con gli stati del mondo, ma la questione di come esse possano "significare" qualcosa per il sistema stesso. Un termostato ha stati interni correlati con la temperatura dell'ambiente, ma non si direbbe che "rappresenta" la temperatura nello stesso senso in cui una mente la rappresenta. La correlazione causale e la funzione sistemica non sono, da sole, sufficienti a costituire l'intentionalità genuina.

Il framework del cervello predittivo fornisce una risposta più sofisticata del termostato: le rappresentazioni del modello generativo non sono semplici correlati di

stati del mondo, ma strutture che codificano relazioni causali tra stati del mondo e input sensoriali, e che possono essere composte in modi flessibili e generativi. Tuttavia, nemmeno questo sembra sufficiente per rispondere alla domanda sull'intentionalità nella sua accezione più forte — perché si può sempre chiedere: per chi sono queste strutture significative? Qual è il soggetto per cui il modello generativo "riguarda" il mondo?

Questa domanda conduce direttamente al problema più arduo, quello della coscienza e della soggettività, che sarà al centro della terza parte. Ma è possibile fin da qui identificare una tensione strutturale nel paradigma: il framework variazionale è una teoria in terza persona — descrive strutture computazionali che possono essere studiate dall'esterno, misurate, simulate. L'esperienza percettiva è un fenomeno in prima persona — ha un carattere qualitativo che non si esaurisce nella sua descrizione funzionale. Colmare questa tensione richiederà un passaggio concettuale che il formalismo variazionale, da solo, non può compiere.

3. Il Soggetto come Minimo di Energia Libera — Markov Blanket, Emergenza e il Problema della Coscienza

3.1 Il confine tra sé e mondo: la nozione di coperta di Markov

Il framework variazionale, nella sua forma più ambiziosa, non si limita a descrivere come il cervello elabori l'informazione sensoriale. Esso pretende di offrire un principio unificante che spieghi l'esistenza stessa dei sistemi cognitivi — e, in ultima analisi, dei sistemi biologici in quanto tali — come conseguenza di un principio di minimizzazione dell'energia libera che opera a scale temporali e spaziali diverse. Per comprendere questa pretesa nella sua portata piena, è necessario introdurre una nozione di origine statistica che Friston ha posto al centro della sua costruzione teorica: la coperta di Markov, o Markov blanket.

Il concetto ha origine nella teoria delle reti bayesiane, dove fu introdotto da Judea Pearl. In una rete bayesiana — un grafo aciclico orientato in cui i nodi rappresentano variabili aleatorie e gli archi rappresentano dipendenze causali — la coperta di Markov di un nodo X è l'insieme minimo di nodi che rende X statisticamente indipendente da tutti gli altri nodi della rete, condizionatamente ai nodi della coperta stessa. Formalmente, indicando con $mb(X)$ la coperta di Markov di X e con R il resto della rete:

$$P(X \mid mb(X), R) = P(X \mid mb(X))$$

La coperta di Markov di X comprende i suoi genitori diretti nel grafo — le variabili che influenzano causalmente X — i suoi figli diretti, e i co-genitori dei suoi figli — le altre variabili che influenzano causalmente i figli di X . È la coperta minimale che "scherma" X dall'influenza diretta del resto della rete.

Friston propone di leggere questa struttura statistica come la base formale della distinzione tra un sistema e il suo ambiente. Un sistema fisico — un organismo biologico, un cervello, una cellula — può essere identificato con un insieme di variabili interne μ che sono separate dall'ambiente esterno η attraverso una coperta di Markov b . Le variabili della coperta si dividono a loro volta in due classi: gli stati sensoriali s , che sono influenzati dall'ambiente esterno ma non dalle variabili interne in modo diretto, e gli stati attuativi a , che influenzano l'ambiente esterno e sono influenzati dalle variabili interne. Schematicamente:

$$\eta \rightarrow s \rightarrow \mu \rightarrow a \rightarrow \eta$$

dove le frecce indicano le influenze causali dirette, e la coperta $b = \{s, a\}$ costituisce il confine tra il sistema interno μ e l'ambiente esterno η . La struttura è ciclica: l'ambiente influenza il sistema attraverso gli stati sensoriali, il sistema risponde attraverso gli stati attuativi, e le azioni modificano l'ambiente che genera i futuri stati sensoriali — un loop di percezione-azione continuo.

3.2 L'esistenza come minimizzazione: auto-organizzazione e identità statistica

La tesi più audace di Friston è che la mera esistenza di una coperta di Markov stabile — la persistenza nel tempo di un confine statistico tra il sistema e il suo ambiente — sia equivalente, in condizioni di non-equilibrio, alla minimizzazione dell'energia libera. L'argomentazione procede come segue.

Un sistema fisico lontano dall'equilibrio termodinamico tende verso uno stato stazionario di non-equilibrio — una distribuzione stazionaria sugli stati del sistema che non è quella di massima entropia ma riflette la struttura delle sue interazioni con l'ambiente. Questa distribuzione stazionaria può essere formalizzata come la distribuzione ergodica $p^*(\mu, s, a, \eta)$ del sistema. L'energia libera variazionale di un certo stato del sistema rispetto a questa distribuzione stazionaria è:

$$\begin{aligned} F &= E_Q[\ln Q(\eta \mid b)] - E_Q[\ln p^*(\eta \mid b)] \\ &= D_{KL}(Q(\eta \mid b) \parallel p^*(\eta \mid b)) \end{aligned}$$

dove $Q(\eta \mid b)$ è la distribuzione di probabilità sugli stati dell'ambiente inferita dal sistema a partire dagli stati della coperta b , e $p^*(\eta \mid b)$ è la distribuzione stazionaria degli stati ambientali condizionatamente agli stati della coperta. Poiché $D_{KL} \geq 0$, $F \geq 0$, con uguaglianza solo quando la distribuzione interna coincide con quella stazionaria. Un sistema che persiste nel tempo — che mantiene la propria integrità strutturale contro le perturbazioni dell'ambiente — è precisamente un sistema che

rimane in prossimità della sua distribuzione stazionaria, il che significa che esso minimizza l'energia libera variazionale in modo continuo.

Questa argomentazione ha una struttura circolare che è importante riconoscere: la distribuzione stazionaria p^* è definita rispetto alle dinamiche del sistema, e la minimizzazione di F rispetto a p^* non è un principio normativo esterno ma una tautologia che descrive la condizione di sopravvivenza del sistema. Il sistema che persiste è quello che minimizza l'energia libera rispetto alla propria distribuzione stazionaria — per definizione, poiché la distribuzione stazionaria è precisamente la distribuzione degli stati compatibili con la sopravvivenza del sistema. Questa circolarità sarà oggetto di critica nella quarta parte; per il momento, si noti che essa non rende il framework privo di contenuto empirico, ma implica che le sue predizioni più specifiche richiedono una specificazione indipendente della distribuzione stazionaria.

3.3 Dall'identità statistica alla soggettività: un salto concettuale

Ammettendo che la struttura della coperta di Markov fornisca un criterio formale per l'identità di un sistema fisico — per la distinzione tra ciò che è "interno" e ciò che è "esterno" — si pone la questione fondamentale: questo criterio è sufficiente a dar conto dell'identità soggettiva? L'io percipiente — il soggetto dell'esperienza — è riducibile alla struttura statistica della coperta di Markov di un sistema neurale?

La risposta che si propone qui è negativa, ma per ragioni che meritano un esame articolato. La coperta di Markov fornisce un criterio formale di individuazione — essa definisce un confine tra il sistema e il suo ambiente in termini di dipendenze statistiche. Ma il soggetto dell'esperienza non è soltanto un sistema individuato in senso statistico: esso è un sistema per cui c'è qualcosa che è essere quel sistema, nel senso della celebre formula di Thomas Nagel. La distinzione tra esterno e interno che la coperta di Markov formalizza è una distinzione epistemica e computazionale, non una distinzione fenomenologica.

Consideriamo un semplice esempio. Un termostato, pur privo di qualunque interesse filosofico, può essere descritto come dotato di una coperta di Markov che lo individua come sistema distinto dall'ambiente termico. Gli stati interni del termostato sono condizionatamente indipendenti dalla temperatura esterna dato il valore del sensore — il sensore è parte della coperta sensoriale del termostato. Nessuno sarebbe tentato di attribuire al termostato un'esperienza soggettiva della temperatura. La struttura della coperta di Markov è quindi necessaria ma non sufficiente per la soggettività.

Il framework variazionale, nella sua versione più raffinata, è consapevole di questa insufficienza e propone di colmarla attraverso la nozione di inferenza attiva: un sistema ha qualcosa di più simile a un'esperienza soggettiva nella misura in cui i suoi stati interni non si limitano a correlare passivamente con gli stati dell'ambiente, ma costruiscono attivamente modelli di quell'ambiente — modelli dotati di struttura causale e predittiva — e li usano per guidare le proprie azioni. Ma anche questa proposta, si argomenterà, non è sufficiente: la costruzione di modelli causali

dell'ambiente può avvenire in sistemi che non hanno alcuna esperienza soggettiva manifesta.

3.4 Il problema difficile della coscienza e il suo rapporto con il predictive coding

David Chalmers, nel 1995, ha articolato con precisione la distinzione tra i cosiddetti "problemi facili" della coscienza e il "problema difficile". I problemi facili riguardano le funzioni cognitive della coscienza — la discriminazione degli stimoli, l'integrazione dell'informazione, il controllo del comportamento, la reportabilità verbale degli stati interni — e sono "facili" non nel senso che siano privi di complessità tecnica, ma nel senso che, in linea di principio, possono essere affrontati attraverso l'analisi funzionale e computazionale. Il problema difficile riguarda invece il perché ci sia esperienza soggettiva che accompagna queste funzioni — perché il cervello che elabora l'informazione visiva non lo faccia "al buio", senza che vi sia alcun carattere qualitativo associato all'elaborazione.

Il framework del predictive coding è straordinariamente eloquente riguardo ai problemi facili. La discriminazione percettiva si spiega come la risoluzione di ambiguità nel modello generativo attraverso la minimizzazione dell'energia libera. L'integrazione dell'informazione si spiega come la coerenza tra rappresentazioni a livelli diversi della gerarchia, garantita dalla propagazione degli errori di predizione. Il controllo del comportamento si spiega come inferenza attiva. La reportabilità verbale si spiega come l'accesso dei livelli superiori della gerarchia agli errori di predizione dei livelli inferiori, che permette di incorporare le rappresentazioni percettive nel modello generativo di ordine superiore che governa la produzione del linguaggio. Per tutte queste funzioni, il framework offre conti meccanicistici precisi e verificabili.

Ma il problema difficile rimane intatto. Che cosa rende l'aggiornamento bayesiano delle rappresentazioni un'esperienza soggettiva? Perché la soppressione degli errori di predizione è accompagnata da un vissuto fenomenologico — dal "come è" percepire il rosso, sentire il dolore, ascoltare il silenzio dopo una lunga nota musicale? Il framework variazionale non fornisce, da solo, nessuna risposta a questa domanda. E non è chiaro che possa fornirla senza un supplemento concettuale che vada oltre la struttura dell'inferenza.

3.5 Tentativi di risposta: dall'integrazione dell'informazione alla coscienza predittiva

Due linee di risposta meritano esame, pur con le loro difficoltà interne. La prima è la teoria dell'integrazione dell'informazione di Giulio Tononi — la cosiddetta IIT — che identifica la coscienza con una misura quantitativa dell'integrazione causale dell'informazione, indicata con il simbolo Φ (phi). La seconda è il programma di una "scienza della coscienza" basata sul predictive coding sviluppato, in forme diverse, da Jakob Hohwy, Andy Clark e Anil Seth.

La IIT definisce Φ come la quantità di informazione generata da un sistema come un tutto, al di là dell'informazione generata dalle sue parti considerate indipendentemente. Formalmente, per un sistema S con stato corrente x e stato passato x_0 , Φ è definito come:

$$\Phi = \min_P [\varphi(S, P)]$$

dove il minimo è preso sull'insieme delle partizioni bipartite P del sistema, e $\varphi(S, P)$ è la perdita di informazione causale che si produce quando si "taglia" il sistema secondo la partizione P — misurata attraverso la divergenza di Earth Mover tra la distribuzione degli stati del sistema intero e il prodotto delle distribuzioni degli stati delle due parti:

$$\varphi(S, P) = d_{EM}(p(x | x_0) \parallel p(x^A | x_0^A) \times p(x^B | x_0^B))$$

con A e B che indicano le due parti della partizione. Un sistema è cosciente se e solo se $\Phi > 0$, e il livello della sua coscienza è proporzionale a Φ .

L'attrattiva della IIT sta nella sua precisione formale e nella sua capacità di distinguere tra sistemi funzionalmente equivalenti con diversi gradi di integrazione causale — il cervello umano avrebbe un Φ molto elevato, mentre un insieme di neuroni scollegati o un sistema feedforward puro avrebbero $\Phi = 0$ o basso. Ma la IIT soffre di difficoltà serie: il calcolo esatto di Φ è computazionalmente intrattabile per sistemi di dimensioni neurologicamente realistiche; la misura di Earth Mover dipende dalla scelta di una metrica sullo spazio degli stati che è in larga misura arbitraria; e soprattutto, la IIT implica che certi sistemi artificiali molto semplici con alta integrazione abbiano un grado di coscienza superiore a quello di sistemi biologici molto più complessi, un risultato che molti trovano controintuitivo.

La seconda linea di risposta — quella che potrebbe chiamarsi "coscienza predittiva" — non identifica la coscienza con una misura quantitativa ma la associa a una caratteristica qualitativa del modo in cui il modello generativo elabora l'informazione. Nella versione di Anil Seth, il sé cosciente è esso stesso una costruzione percettiva — una "ipotesi bestiale" del cervello sulla causa delle sue stesse afferenze sensoriali, intero-cettive e propriocettive. L'esperienza soggettiva del sé — il senso di essere un agente corporeo situato nel tempo e nello spazio — è il contenuto di un modello generativo del sistema su se stesso: un'inferenza sulle cause degli stati interni tanto quanto la percezione visiva è un'inferenza sulle cause degli stati retinali.

Questa proposta ha il merito di estendere il framework predittivo alla sfera della coscienza di sé in modo coerente con la sua architettura generale. Ma essa trasforma, senza risolvere, il problema difficile: invece di chiedersi perché vi sia esperienza soggettiva che accompagna l'elaborazione sensoriale del mondo esterno, ci si trova a chiedersi perché vi sia esperienza soggettiva che accompagna il modello generativo del sé. Il problema regredisce di un livello, ma non scompare.

3.6 Il corpo come prior: interocettività e affetto

Un contributo genuinamente originale che il framework predittivo porta alla filosofia della mente riguarda il ruolo del corpo — e in particolare degli stati viscerali e interoceptivi — nella costituzione dell'esperienza cosciente. La tradizione cartesiana ha tendenzialmente opposto la mente al corpo, concependo la soggettività come qualcosa che si sovrappone a un substrato corporeo ma non ne è costituita in modo essenziale. Il framework predittivo suggerisce invece che il corpo — nel senso delle sue afferenze interoceptive, delle segnalazioni vagali, degli stati ormonali — svolga il ruolo di distribuzione a priori fondamentale su cui si costruisce l'intera gerarchia delle inferenze percettive.

L'interocezione — la percezione degli stati interni del corpo, distinta dalla propriocezione che riguarda la postura e il movimento — fornisce un segnale di cruciale importanza per la regolazione omeostatica. Dalla prospettiva bayesiana, gli stati viscerali comunicano al sistema centrale informazioni sulla distanza attuale dagli stati fisiologici ottimali — dal set point dell'omeostasi. Il disagio fisico è un errore di predizione interoceptivo: una discrepanza tra lo stato corporeo attuale e lo stato corporeo che il modello generativo si aspetterebbe in condizioni di benessere. La fame è l'errore di predizione metabolico; il dolore è l'errore di predizione tissutale; l'ansia potrebbe essere interpretata come un errore di predizione interoceptivo di natura più diffusa, una segnalazione di instabilità sistemica che precede qualsiasi identificazione di una causa esterna specifica.

In questa prospettiva, le emozioni — nella loro accezione più elementare, pre-cognitiva — non sono interpretazioni cognitive di stati corporei (come voleva la teoria di James-Lange), né sono stati puramente mentali che si esprimono nel corpo (come implica una certa lettura del cognitivismo), ma sono esse stesse errori di predizione interoceptivi che il sistema tenta di sopprimere attraverso azioni regolatorie. L'affetto — la dimensione di piacere/dolore, attivazione/inibizione che permea ogni esperienza — corrisponde alla valenza e all'intensità degli errori di predizione interoceptivi residui.

Questa proposta ha conseguenze filosofiche notevoli. Essa implica che la coscienza non può essere compresa facendo astrazione dal corpo: il carattere fenomenologico dell'esperienza — il suo "tono affettivo" di fondo — è radicato negli stati viscerali e nelle loro deviazioni dalle aspettative omeostatiche. Un'intelligenza artificiale che mancasse di interocezione — di un equivalente funzionale degli stati corporei e delle loro dinamiche — non soltanto mancherebbe di emozioni nel senso ordinario, ma mancherebbe dell'intera dimensione affettiva che, secondo questa prospettiva, caratterizza il tessuto dell'esperienza cosciente. Non è una questione di mancare di un "aggiunta" emotiva a un'elaborazione cognitiva altrimenti completa: è una questione di mancare della struttura fondamentale che rende l'elaborazione cognitiva soggettivamente rilevante.

3.7 Il sé minimo e il sé narrativo: livelli della soggettività

Una distinzione che il framework predittivo permette di articolare con nuova precisione è quella tra il sé minimo — il senso pre-riflessivo di essere un soggetto

corporeo situato nel tempo e nello spazio, che Husserl chiamava *Urpräsenz* e Damasio ha esplorato nella sua nozione di "proto-sé" — e il sé narrativo — la costruzione autobiografica e riflessiva che organizza l'esperienza in una storia coerente con un protagonista identificabile attraverso il tempo.

Il sé minimo corrisponde, nel framework predittivo, al livello più basso e più fondamentale della gerarchia del modello generativo: il modello che rappresenta il corpo come agente situato nello spazio tridimensionale, dotato di un campo visivo, di una postura, di stati viscerali. Questo modello è aggiornato continuamente e in tempo reale attraverso l'integrazione di segnali propriocettivi, vestibolari, visivi e interoceptivi. Esso è il substrato dell'esperienza di essere qui-e-ora, di avere un punto di vista percettivo, di essere l'origine delle proprie azioni.

Il sé narrativo corrisponde invece ai livelli superiori della gerarchia: al modello del proprio carattere, delle proprie relazioni, della propria storia passata e delle proprie aspettative future. Questo modello è costruito attraverso l'integrazione nel tempo di grandi quantità di esperienze passate, mediata dalla memoria autobiografica e dalla narrazione linguistica. È il sé che risponde alla domanda "chi sei?" con una storia — un racconto di esperienze, relazioni, scelte, valori.

La distinzione tra questi due livelli ha implicazioni cliniche importanti. Alcune condizioni psicopatologiche possono essere interpretate come perturbazioni della coperta di Markov al livello del sé minimo — il senso di confine tra sé e mondo si attenua o si dissolve, come accade in certi stati dissociativi, nella psicosi, o in alcune esperienze di derivante dalla pratica meditativa intensa. Altre condizioni perturbano prevalentemente il sé narrativo — la depressione, in cui la storia di sé diventa dominata da aspettative a priori negative e resistenti all'aggiornamento; certi disturbi di personalità, in cui il modello narrativo del sé è costruito intorno a schemi relazionali rigidi che strutturano ogni nuova esperienza in modi difensivi.

4. Critiche, Limiti e Circoli Viziosi del Paradigma

4.1 La tentazione dell'onniscienza

Ogni grande teoria unificante nella scienza porta con sé una tentazione caratteristica: quella di spiegare tutto. Quando il principio esplicativo è abbastanza astratto e abbastanza flessibile, esso rischia di diventare irrefutabile non per la sua verità ma per la sua vaghezza — capace di sussumere qualunque fenomeno sotto la propria struttura, al prezzo di perdere il contenuto empirico che rende una teoria genuinamente informativa. Il framework dell'energia libera variazionale non è immune da questa tentazione, e una critica onesta del paradigma deve affrontarla senza indulgenza.

Il sospetto di onnicomprensività si manifesta già nella struttura dell'argomentazione principale: si dimostra che qualunque sistema che persista nel tempo — qualunque sistema che abbia una distribuzione stazionaria e mantenga la propria integrità strutturale — può essere descritto come se minimizzasse l'energia libera variazionale. Questo "come se" è cruciale. L'affermazione è vera in quanto tautologia: la distribuzione stazionaria è per definizione l'insieme degli stati che il sistema occupa con probabilità stazionaria, e la distanza da quella distribuzione è per definizione l'energia libera variazionale del sistema rispetto a se stesso. Affermare che i sistemi viventi minimizzano l'energia libera è quindi, in questa lettura, una riformulazione della definizione di sistema stazionario, non una scoperta empirica indipendente.

La risposta standard a questa critica è che il framework guadagna contenuto empirico specifico quando si specifica la forma del modello generativo per un sistema particolare. Le previsioni del predictive coding per il sistema visivo — specifiche distribuzioni di attivazione nei neuroni corticali, specifici pattern di propagazione degli errori di predizione — sono empiricamente verificabili e falsificabili. Il framework generale non è tautologico, ma l'istanza specifica è dotata di contenuto. Questa risposta è corretta, ma implica uno spostamento importante: il principio generale dell'energia libera come principio biologico unificante non è direttamente verificabile, e la sua difesa dipende interamente dalla verifica delle sue istanze specifiche — un programma empirico di vasto respiro che è tuttora in corso.

4.2 Il circolo dell'active inference

Una difficoltà strutturale più specifica riguarda la nozione di active inference, la proposta che l'azione sia anch'essa un meccanismo di minimizzazione dell'energia libera attraverso la modificazione degli input sensoriali piuttosto che delle rappresentazioni interne.

Il problema nasce dal modo in cui vengono generate le predizioni motorie. Nel framework standard del predictive coding, l'azione riduce gli errori di predizione propriocettivi: il sistema predice un certo stato propriocettivo — corrispondente alla posizione desiderata dell'arto — e l'errore di predizione tra la posizione attuale e quella predetta viene risolto attraverso il movimento, che porta la posizione attuale in accordo con quella predetta. Ma questo schema implica che le predizioni propriocettive siano generate da livelli superiori della gerarchia che rappresentano le intenzioni motorie.

La domanda cruciale diventa: come vengono generate queste intenzioni motorie di ordine superiore? Se anch'esse sono il prodotto di inferenze variazionali a livelli ancora superiori della gerarchia, allora si apre una regressione: ogni livello di intenzione motoria è generato da un livello superiore di aspettative, e così via. Questa regressione si arresta naturalmente se esiste un livello più alto della gerarchia in cui le aspettative sono fissate da parametri che non sono a loro volta il prodotto di inferenza in tempo reale — per esempio, da pulsioni omeostatiche o da valori profondamente radicati. Ma allora questi livelli apicali della gerarchia svolgono un ruolo normativo che non può essere derivato dalla struttura dell'inferenza variazionale stessa: essi fissano i fini verso

cui l'inferenza è orientata, ma non sono essi stessi il prodotto di quella stessa inferenza.

Questa struttura rivela una tensione tra la descrizione del framework come puramente computazionale-descrittivo — il cervello minimizza l'energia libera, punto — e la necessità di introdurre componenti normative che determinino verso quale distribuzione stazionaria il sistema tende. Il framework non può stabilire da solo verso quale fine il sistema deve tendere: può soltanto descrivere il processo attraverso cui il sistema persegue i fini che già possiede. La questione dell'origine e della struttura di questi fini — le pulsioni primarie, i valori fondamentali, gli scopi ultimi — rimane esterna alla struttura dell'inferenza variazionale.

4.3 Il problema dell'approssimazione di campo medio: conseguenze biologiche

Sul piano strettamente tecnico, l'approssimazione di campo medio — l'ipotesi che la distribuzione approssimante Q si fattorizzi come prodotto di distribuzioni indipendenti su ciascuna variabile latente — è la principale fonte di limitazioni computazionali del framework quando viene applicato a sistemi neurologici realistici.

In un sistema gerarchico con L livelli, ciascuno dei quali è caratterizzato da vettori di stato di dimensione n , il numero di parametri necessari a rappresentare le correlazioni tra tutti i livelli cresce come $O(n^2 L^2)$. L'approssimazione di campo medio riduce questo numero a $O(nL)$ — un guadagno computazionale enorme, ma al prezzo di ignorare tutte le correlazioni tra variabili latenti a livelli diversi della gerarchia. Formalmente, se la vera distribuzione posteriore è $P(s^1, s^2, \dots, s^L \mid o)$, l'approssimazione di campo medio sostituisce:

$$Q(s^1, s^2, \dots, s^L) = \prod_l Q_l(s^l)$$

La divergenza tra la vera posteriore e questa approssimazione fattorizzata può essere misurata dalla mutual information tra i livelli nella vera posteriore:

$$D_{KL}(Q \parallel P) \geq \sum_{l \neq l'} I_P(s^l; s^{l'})$$

dove $I_P(s^l; s^{l'})$ è la mutua informazione tra i livelli l e l' nella vera distribuzione posteriore. Quando questa mutua informazione è significativa — cioè quando ci sono dipendenze forti tra variabili a livelli diversi — l'approssimazione di campo medio introduce errori sistematici nella stima della posteriore, producendo distribuzioni eccessivamente concentrate e stime dei parametri distorte.

Biologicamente, questo si traduce in una rigidità della percezione che potrebbe avere correlati clinici. Un sistema che non tiene conto delle correlazioni tra livelli della gerarchia percettiva sarà meno capace di integrare flessibilmente informazioni provenienti da scale di elaborazione diverse — per esempio, di integrare caratteristiche locali e struttura globale nella percezione degli oggetti. Il fatto che certe condizioni del

neurosviluppo, come lo spettro dell'autismo, siano associate a una percezione atipicamente orientata alle caratteristiche locali piuttosto che alle strutture globali ha suggerito ad alcuni ricercatori — in particolare Pellicano e Burr, e successivamente van Boxtel e colleghi — che esse possano riflettere anomalie nel bilanciamento tra priori e verosimiglianze, o più specificamente nella struttura delle approssimazioni utilizzate dal sistema per rendere trattabile il calcolo dell'inferenza. Questa è una proposta empiricamente fertilissima, ma richiede di essere valutata con cautela: la complessità del fenotipo autistico non si lascia catturare da una singola perturbazione computazionale.

4.4 La questione del tempo: processi dinamici e inferenza su sequenze temporali

Un'altra limitazione significativa del framework, nella sua versione originaria, è la difficoltà di trattare in modo naturale la struttura temporale dell'esperienza percettiva. La formulazione standard del predictive coding tratta le osservazioni sensoriali come eventi istantanei e aggiorna le rappresentazioni in risposta a ciascuna osservazione in modo sostanzialmente asincrono. Questa astrazione è conveniente matematicamente ma biologicamente irrealistica: il cervello non elabora istantanee discrete del mondo ma flussi continui di informazione sensoriale, e la struttura temporale di questi flussi — la durata, il ritmo, la causalità temporale tra eventi — è essa stessa una proprietà del mondo che deve essere rappresentata e inferita.

La teoria delle traiettorie generalizzate, proposta da Friston come estensione del framework standard, tenta di affrontare questo problema rappresentando gli stati del mondo non come punti in uno spazio di stati ma come traiettorie continue nello spazio degli stati, caratterizzate non soltanto dalla loro posizione μ ma anche dalle loro derivate temporali di ordine arbitrario μ' , μ'' , ..., $\mu^{(n)}$ — un embedding dello spazio degli stati nei cosiddetti stati generalizzati o "spazio dei movimenti". Formalmente, per uno stato del mondo che segue una dinamica deterministica perturbata da rumore, la distribuzione sugli stati generalizzati al tempo t è:

$$p(\tilde{\mu} | t) = p(\mu, \mu', \mu'', \dots | t)$$

dove $\tilde{\mu}$ denota il vettore di stati generalizzati. L'energia libera variazionale sui stati generalizzati è:

$$\tilde{F} = D_{\text{KL}}(Q(\tilde{\mu}) \parallel P(\tilde{\mu} | \tilde{o}))$$

con $\tilde{o}$ che indica le osservazioni generalizzate — anch'esse rappresentate con le loro derivate temporali. La minimizzazione di $\tilde{F}$ rispetto al vettore di stati generalizzati porta a equazioni del moto che descrivono l'aggiornamento continuo delle rappresentazioni in risposta al flusso sensoriale, tenendo conto della struttura dinamica dell'elaborazione.

Questo formalismo è matematicamente elegante ma introduce una nuova difficoltà: la scelta dell'ordine n a cui troncare la sequenza delle derivate generalizzate è in larga misura arbitraria, e le conseguenze pratiche di questa scelta per i comportamenti biologici sono poco comprese. Per ordini bassi, il sistema rappresenta soltanto la posizione e la velocità degli stati — una descrizione della dinamica locale. Per ordini elevati, il sistema rappresenta strutture dinamiche sempre più complesse, ma a costo di un aumento del costo computazionale e di una crescente sensibilità all'identificabilità del modello. Non esiste ad oggi un principio derivato dall'interno del framework che stabilisca quale ordine sia ottimale per quale sistema biologico.

4.5 La misurabilità del modello generativo: il problema dell'identificabilità

Una delle sfide empiriche più serie per il framework riguarda la misurabilità del modello generativo. La struttura interna del modello generativo — le distribuzioni a priori $P(s)$, la funzione di verosimiglianza $P(o | s)$, la struttura gerarchica — non è direttamente osservabile. Essa deve essere inferita dal comportamento del sistema o da misure indirette dell'attività neurale, attraverso procedure di fitting di modelli che comportano inevitabilmente scelte di design difficili da giustificare in modo indipendente.

Il problema dell'identificabilità si manifesta con particolare acutezza quando si cerca di distinguere tra due possibili spiegazioni di un dato comportamento percettivo anomalo: una distribuzione a priori alterata $P(s)$ versus una funzione di verosimiglianza alterata $P(o | s)$ versus un meccanismo di minimizzazione alterato. In termini formali, per un insieme di osservazioni comportamentali B , si possono costruire in generale molteplici modelli $M_1, M_2, \dots$ con diversi modelli generativi interni che producono le stesse previsioni comportamentali:

$$P(B | M_1) = P(B | M_2)$$

Questa degenerazione modello-comportamento implica che il comportamento osservato, da solo, non è sufficiente a identificare univocamente il modello generativo del sistema. Si devono integrare misure comportamentali con misure neurali — per esempio, il profilo di attivazione di aree corticali specifiche durante compiti percettivi — ma la relazione tra le strutture computazionali del modello e le strutture neurali osservabili è anch'essa non banale e soggetta a ipotesi di fondo non sempre esplicite.

Questa difficoltà non è specifica del framework variazionale — è una manifestazione generale del problema inverso in neuroscienza cognitiva — ma è particolarmente acuta in un paradigma che pretende di fornire un account computazionale completo della cognizione, poiché le sue previsioni più specifiche dipendono da parametri del modello generativo che sono difficili o impossibili da misurare in modo indipendente.

4.6 La critica ecologica: Gibson e il problema delle affordanze

Una critica di natura diversa, e a mio avviso filosoficamente più profonda di quanto sia spesso riconosciuto, proviene dalla tradizione della psicologia ecologica inaugurata da James Gibson. Gibson ha sostenuto che la percezione non è un processo di inferenza sulle cause nascoste delle sensazioni, ma un processo di rilevamento diretto delle affordanze — delle possibilità di azione che l'ambiente offre all'organismo. Nella prospettiva gibsoniana, il sistema percettivo è accordato direttamente alle strutture informazionali presenti nel flusso ottico e negli altri array energetici, senza che sia necessario postulare modelli interni di ordine superiore.

La critica gibsoniana al framework del cervello predittivo può essere articolata in tre versioni di crescente profondità. La versione debole sostiene semplicemente che il cervello predittivo descrive la percezione a un livello di analisi troppo alto — computazionale, nel senso di Marr — senza specificare i meccanismi algoritmici e implementativi concreti, che potrebbero essere molto diversi da quelli che il framework suggerisce. Questa versione è facilmente respinta: il framework è esplicitamente una teoria computazionale, e la sua integrazione con descrizioni algoritmiche e implementative è un programma aperto.

La versione media sostiene che il framework del cervello predittivo sottovaluta sistematicamente il ruolo dell'informazione direttamente disponibile nell'array di stimolazione — quella che Gibson chiamava "informazione di ordine superiore" presente nella struttura invariante del flusso ottico — e sovrastima il contributo dei modelli interni. In certi domini della percezione — il controllo del movimento, la navigazione visiva, la percezione del moto biologico — l'informazione disponibile nello stimolo è sufficiente a determinare il comportamento senza che sia necessario invocare modelli generativi complessi. Questa versione è più sostanziosa e ha ricevuto supporto empirico in certi contesti, ma non contraddice in linea di principio il framework bayesiano: un sistema bayesiano con distribuzioni a priori molto diffuse — poco informative — si comporta in modo simile a un sistema che elabora direttamente l'informazione sensoriale senza forti aspettative.

La versione forte, che è la più interessante, sostiene che il concetto stesso di "modello generativo interno" è fuorviante perché reifica una distinzione tra sistema e mondo che non è ontologicamente primaria. Nella prospettiva di Gibson — e più ancora nella sue estensioni enattiviste sviluppate da Maturana, Varela e Thompson — il sistema cognitivo non è il cervello isolato che elabora input sensoriali provenienti da un mondo esterno, ma il circolo sensomotorio organismo-ambiente considerato nella sua integralità. Le affordances non sono proprietà dell'ambiente che il cervello deve inferire partendo da segnali sensoriali depauperati: sono relazioni tra le capacità dell'organismo e le strutture dell'ambiente, direttamente percepibili da un organismo adeguatamente sintonizzato.

Il framework variazionale, nella sua versione più ambiziosa — quella che include l'active inference e il ciclo percezione-azione — si avvicina a questa prospettiva: il sistema che minimizza l'energia libera non è il cervello in isolamento ma l'organismo nel suo ciclo di interazione con l'ambiente. Ma c'è una differenza concettuale che rimane: il framework variazionale mantiene la distinzione tra variabili interne e variabili

esterne — μ e η , separati dalla coperta di Markov — mentre la prospettiva enattivista forte nega che questa distinzione sia primaria. Questa non è una differenza meramente terminologica: essa ha conseguenze per la questione su dove si collochi il "sito" della cognizione, e per il problema dell'intentionalità originaria discusso nella terza parte.

4.7 Il problema dell'apprendimento: come si acquisisce il modello generativo?

Il framework nella sua versione standard tratta il modello generativo come dato — descrive come un sistema con un certo modello generativo aggiorna le sue credenze e guida le sue azioni, ma lascia largamente non specificato come quel modello si sia sviluppato in primo luogo. La questione dell'apprendimento — dell'acquisizione del modello generativo attraverso l'esperienza — è affrontata nella letteratura attraverso la distinzione tra inferenza percettiva, che opera su scale temporali rapide (millisecondi-secondi), e apprendimento, che opera su scale temporali lente (minuti-anni) attraverso la modifica dei parametri del modello generativo stesso.

Formalmente, l'apprendimento dei parametri θ del modello generativo procede attraverso la minimizzazione dell'energia libera rispetto a θ , piuttosto che rispetto alle rappresentazioni μ :

$$\Delta\theta \propto -\partial F/\partial\theta$$

Nel caso di un modello generativo con struttura gaussiana, questo gradiente rispetto ai parametri porta a regole di aggiornamento che sono analoghe formalmente alle regole di Hebb per il potenziamento sinaptico — un risultato suggestivo che connette il framework bayesiano alla biofisica della plasticità sinaptica.

Tuttavia, questo schema di apprendimento presenta problemi seri quando si considera l'acquisizione di strutture di ordine superiore — categorie concettuali, relazioni causali complesse, strutture linguistiche. L'apprendimento bayesiano gerarchico richiede che le distribuzioni a priori ai livelli superiori della gerarchia vengano acquisite attraverso l'esperienza accumulata, il che implica che nelle prime fasi dell'apprendimento il sistema abbia distribuzioni a priori molto diffuse e poco informative — un bambino che apprende la struttura degli oggetti visivi non può avere priori forti su quella struttura finché non ne ha accumulata sufficiente esperienza. Ma in assenza di priori forti, l'aggiornamento bayesiano converge molto lentamente, e il sistema è vulnerabile a ogni perturbazione del segnale sensoriale.

Questo è noto come il problema dello "spazio delle ipotesi": il framework variazionale può descrivere come si aggiorna la distribuzione su un insieme di ipotesi dato, ma lascia non specificato come viene definito quello spazio di ipotesi in primo luogo. La struttura dello spazio delle ipotesi — quali categorie di stati del mondo vengono considerate, quali strutture causali vengono rappresentate, quali regolarità vengono estratte — è essa stessa un prodotto dell'apprendimento, ma è un apprendimento di ordine superiore che richiede meccanismi diversi dall'aggiornamento bayesiano

standard. Il framework bayesiano gerarchico — la cosiddetta "macchina di Bayes su Bayes" che impara prior su prior — è il tentativo più elaborato di affrontare questa difficoltà, ma produce una regressione che si arresta soltanto con l'introduzione di prior non appresi, biologicamente determinati — la struttura della "conoscenza innata" nel senso di Chomsky o di Spelke.

4.8 Verso una valutazione complessiva: ciò che rimane

Dopo aver percorso le critiche principali, è doveroso formulare un giudizio complessivo che non cada né nell'entusiasmo acritico né nel rigetto pregiudiziale. Il framework dell'energia libera variazionale e del predictive coding è, a parere di chi scrive, uno dei contributi più fertili e strutturalmente solidi che la ricerca neuroscientifica abbia prodotto negli ultimi decenni. La sua forza risiede precisamente nell'integrare sotto un'unica struttura formale fenomeni che la ricerca tradizionale studiava in compartimenti separati: la percezione, l'attenzione, l'azione, l'apprendimento, la regolazione omeostatica, certi aspetti della psicopatologia. Questa integrazione non è puramente formale — essa genera previsioni specifiche e falsificabili che hanno trovato supporto empirico in molteplici domini.

Le sue debolezze sono altrettanto reali. Il framework è troppo generale per essere direttamente falsificabile nella sua versione più ambiziosa; l'approssimazione di campo medio introduce distorsioni sistematiche la cui rilevanza biologica è ancora poco compresa; il problema del tempo e della struttura dinamica dell'elaborazione rimane parzialmente irrisolto; il problema dell'apprendimento delle strutture di ordine superiore rimanda a un piano di analisi che il framework da solo non riesce a coprire; e il problema della coscienza — il salto dal computazionale al fenomenologico — rimane esattamente dov'era prima che il framework fosse formulato.

Ciò che il framework ha il merito incontestabile di aver chiarito è questo: la mente non è uno specchio del mondo. Essa è un modello del mondo — un modello generativo che costruisce attivamente la propria versione della realtà, la confronta continuamente con i segnali sensoriali disponibili, e li usa non per descrivere il mondo ma per correggere e affinare se stesso. In questa prospettiva, l'errore è più informativo della conferma; il silenzio dell'errore — la soppressione degli scarti tra predizione e realtà — è la forma più alta del conoscere; e l'esperienza di vedere, udire, toccare, sentire è sempre e irriducibilmente il prodotto di un'ipotesi sul mondo che il mondo ha, per il momento, smesso di contraddire.

Il cervello è un'ipotesi che non smette mai di verificarsi. E forse — sebbene su questo punto la certezza rimanga al di là della nostra portata — l'esperienza soggettiva è ciò che accade quando quell'ipotesi si confronta con se stessa.

5. Il Tempo come Struttura del Modello — Causalità, Ritmo e Linguaggio

5.1 La percezione del tempo non è la percezione nel tempo

Esiste una distinzione che la filosofia della mente non può permettersi di trascurare, sebbene la psicologia cognitiva la ignori con sorprendente sistematicità: la distinzione tra percepire nel tempo — il fatto ovvio che ogni atto percettivo si svolge in una successione temporale — e percepire il tempo — il fatto meno ovvio che il tempo stesso sia un oggetto dell'esperienza percettiva, dotato di struttura e contenuto propri. Udire una melodia non è udire una sequenza di note discrete: è cogliere la melodia come unità che si dispiega, in cui ogni nota porta in sé il ricordo delle precedenti e l'attesa delle successive, e in cui il senso dell'insieme non è riducibile alla somma dei contributi istantanei. Questa struttura — che Husserl ha analizzato con incomparabile sottigliezza nelle sue Lezioni sulla fenomenologia della coscienza interna del tempo — è precisamente ciò che il framework del predictive coding standard, nella sua versione originaria, non riesce a catturare.

Il problema è strutturale. Il predictive coding standard tratta le osservazioni sensoriali come variabili discrete, aggiornate a intervalli regolari, e il modello generativo come una macchina che genera distribuzioni su osservazioni istantanee condizionatamente agli stati correnti. Ma l'esperienza temporale — la percezione del ritmo, della durata, della causalità tra eventi — non è riducibile a una sequenza di percezioni istantanee: essa richiede che il sistema mantenga una rappresentazione esplicita della struttura temporale degli eventi, delle loro relazioni di precedenza e conseguenza, della loro velocità e del loro ritmo. La semplice serializzazione di aggiornamenti bayesiani istantanei non produce, da sola, la ricchezza fenomenologica della coscienza del tempo.

5.2 Strutture dinamiche generalizzate: la geometria del movimento

L'estensione del framework alle traiettorie generalizzate, introdotta nella quarta parte in modo schematico, merita qui uno sviluppo che ne metta in luce le conseguenze per la percezione temporale. Il nucleo dell'idea è che il modello generativo non rappresenti soltanto il valore corrente degli stati del mondo — la posizione μ di un oggetto — ma la sua intera traiettoria locale nello spazio degli stati, codificata attraverso le derivate temporali di ordine arbitrario:

$$\tilde{\mu} = (\mu, \mu', \mu'', \mu''', \dots)$$

dove $\mu' = d\mu/dt$ è la velocità, $\mu'' = d^2\mu/dt^2$ è l'accelerazione, e così via. Questo vettore di stati generalizzati vive in uno spazio di dimensione infinita — lo spazio dei germi di traiettorie — che può essere approssimato troncando la sequenza a un ordine finito n , ottenendo uno spazio di dimensione $(n+1) \cdot \dim(\mu)$.

La dinamica di questo sistema è governata da un operatore di derivazione temporale D che agisce sullo spazio degli stati generalizzati secondo lo schema:

$$D \cdot \tilde{\mu} = (\mu', \mu'', \mu''', \dots) \approx A \cdot \tilde{\mu}$$

dove A è la matrice di shift che mappa ogni componente nella sua derivata. Il modello generativo in questo spazio ampliato specifica non soltanto quale sia lo stato più probabile del mondo in un dato momento, ma quale sia la traiettoria più probabile — la curva nello spazio degli stati che meglio predice l'evoluzione futura dello stimolo.

La conseguenza per la percezione del movimento è immediata e verificabile: un sistema che operi sugli stati generalizzati percepirà il movimento non come una sequenza di posizioni discrete ma come una traiettoria continua dotata di struttura intrinseca — velocità, accelerazione, curvatura. Questo è coerente con la psicofisica del movimento visivo, dove i soggetti umani mostrano capacità predittive notevolmente accurate per traiettorie regolari, e dove l'interruzione della continuità di una traiettoria produce un'esperienza percettiva distinta dalla semplice assenza di stimolazione — il fenomeno noto come "motion extrapolation" o estrapolazione del moto.

Ma la struttura degli stati generalizzati permette qualcosa di più ricco della semplice estrapolazione lineare. Se il modello generativo ha imparato la struttura statistica delle traiettorie tipiche in un certo dominio — per esempio, le traiettorie balistiche degli oggetti lanciati, o i pattern di movimento del passo umano — allora esso può predire traiettorie future sulla base di priori costruiti dall'esperienza passata con quella classe di movimenti. La percezione del moto biologico — la capacità di riconoscere immediatamente un essere umano che cammina da un insieme di luci puntiformi attaccate alle articolazioni — si spiegherebbe come il risultato dell'attivazione di un priore sulle traiettorie tipiche del movimento umano che domina rapidamente sull'evidenza sensoriale minima disponibile.

5.3 La coscienza del tempo interno: Husserl incontro a Bayes

Edmund Husserl, nelle sue analisi della coscienza del tempo, ha distinto tre componenti strutturali dell'esperienza temporale: la ritenzione — la persistenza modificata del momento appena trascorso nell'orizzonte del presente vivo — la protoimpressione — l'afferramento del momento presente nella sua immediatezza — e la protensione — l'anticipazione del momento immediatamente futuro. Questi tre momenti non sono separati nel tempo ma sono co-presenti in ogni istante della coscienza: ogni "ora" è intrinsecamente articolato in un appena-stato e un quasi-sarà.

La struttura husserliana della coscienza del tempo non è soltanto una descrizione fenomenologica: è una struttura formale che può essere messa in corrispondenza con l'architettura degli stati generalizzati. La protoimpressione corrisponde al valore corrente μ . La ritenzione corrisponde alla storia recente codificata nelle derivate negative — $\mu(t-\delta)$ per δ piccolo — che il sistema mantiene implicitamente attraverso i suoi stati dinamici. La protensione corrisponde alle predizioni delle derivate positive — $\mu'(t) \cdot \delta, \mu''(t) \cdot \delta^2/2$, ecc. — che proiettano la traiettoria corrente nel futuro prossimo.

Questa corrispondenza non è soltanto analogica. Essa suggerisce che la struttura fenomenologica dell'esperienza temporale — il carattere del presente vivo come esteso, articolato, continuo — possa essere il correlato soggettivo della struttura computazionale degli stati generalizzati. Un sistema che rappresenta la realtà attraverso stati generalizzati di ordine n ha, per così dire, una "finestra temporale" di ampiezza proporzionale a n — una finestra entro cui il passato prossimo e il futuro prossimo sono co-presenti con il momento attuale nella struttura stessa delle sue rappresentazioni. La durata soggettiva del presente vissuto — il "presente specioso" di William James, stimato sperimentalmente tra 2 e 3 secondi — potrebbe corrispondere all'ampiezza di questa finestra nel sistema cognitivo umano.

Questa è una proposta genuinamente originale che, per quanto ne sappia, non è stata sviluppata esplicitamente nella letteratura corrente sul predictive coding. Essa ha il merito di connettere due tradizioni che si sono ignorate a lungo — la fenomenologia husserliana e le neuroscienze computazionali bayesiane — attraverso una struttura formale precisa piuttosto che attraverso analogie vaghe. Ha anche il merito di generare previsioni empiriche: se la durata del presente specioso corrisponde all'ordine degli stati generalizzati mantenuti dal sistema, allora manipolazioni farmacologiche o patologiche che alterino la durata del presente specioso dovrebbero produrre alterazioni prevedibili nell'ordine effettivo degli stati generalizzati misurabile attraverso l'attività neurale. L'espansione del presente vissuto nell'esperienza psilosibinica, o la sua contrazione in certi stati dissociativi, potrebbero essere interpretati in questi termini.

5.4 La percezione della causalità: Hume incontro alla gerarchia predittiva

La percezione della causalità — il fatto che certi eventi siano percepiti come cause di altri, e non semplicemente come eventi concomitanti o successivi — è uno dei fenomeni che più ha affascinato i filosofi dall'epoca di Hume. Hume ha sostenuto con vigore che la causalità non è una proprietà osservabile del mondo ma un'abitudine della mente: non si percepisce mai una connessione necessaria tra causa ed effetto, soltanto una successione regolare, e la mente inferisce la causalità per abitudine associativa quando la successione si ripete abbastanza frequentemente.

Il framework predittivo offre una lettura della percezione della causalità che va oltre l'associazionismo humeano pur mantenendone la struttura anti-realista di fondo. In un modello generativo gerarchico, la causalità è codificata come dipendenza condizionale: A causa B significa che la distribuzione di probabilità su B dato A è diversa dalla distribuzione marginale su B. Formalmente:

$$P(B \mid A) \neq P(B)$$

Ma questa è ancora la nozione statistica di dipendenza, non di causalità in senso direzionale. La struttura causale — il fatto che sia A a causare B e non viceversa, o che

entrambi siano effetti di una causa comune C — richiede un formalismo più ricco, che è offerto dalla teoria dei grafi causali di Pearl.

In un grafo causale, la direzione delle frecce codifica le dipendenze causali in modo che permette di distinguere tra correlazione e causalità: due variabili correlate possono essere legate da una causa comune senza che l'una causi l'altra, e questa distinzione è codificata nella struttura del grafo attraverso i pattern di indipendenza condizionale. Il modello generativo di un sistema cognitivo, in questa prospettiva, è un grafo causale — o una distribuzione su grafi causali possibili — che il sistema mantiene sul mondo, e che gli permette di distinguere tra coincidenze statistiche e relazioni causali genuine.

La percezione della causalità — il fenomeno sperimentato classicamente negli esperimenti di Michotte, dove certi pattern di movimento sono percepiti come "lancio" o "spinta" con carattere fenomenologico di necessità causale — si spiegherebbe come la rapida convergenza del modello generativo verso un grafo causale in cui un oggetto è il generatore del movimento dell'altro. La forza con cui questa percezione causale si impone — la sua resistenza alla conoscenza consapevole della sua natura illusoria in certi display artificiali — è la conseguenza della precisione elevata che il priore sulla causalità meccanica ha nel sistema cognitivo adulto, priore forgiato da anni di esperienza con oggetti fisici che interagiscono per contatto.

5.5 Il linguaggio come modello generativo di ordine superiore

Il linguaggio è il banco di prova più severo per qualunque teoria della cognizione. Esso è lo strumento con cui la mente esprime la propria struttura — il luogo in cui la capacità di modellare il mondo raggiunge la massima articolazione, la massima generalità e la massima dipendenza dalla struttura sociale e culturale. La questione di come il framework predittivo possa dar conto della comprensione e produzione del linguaggio è quindi non soltanto una questione tecnica ma una questione filosofica di prim'ordine.

La proposta di base è relativamente naturale: il linguaggio è un dominio in cui le predizioni del modello generativo operano su sequenze di simboli piuttosto che su segnali sensoriali continui, e la comprensione linguistica è il processo di costruzione di un modello generativo delle strutture simboliche che spieghi la sequenza di input linguistici ricevuti. Questa proposta trova il suo equivalente computazionale nei modelli linguistici probabilistici gerarchici — in particolare nei transformer e nei modelli di lingua di grandi dimensioni — che modellano esplicitamente la distribuzione di probabilità sulle sequenze di token condizionatamente al contesto precedente.

Ma la connessione va più in profondità della semplice analogia strutturale. Nel framework predittivo, la comprensione di una frase è il processo attraverso cui il sistema aggiorna iterativamente la sua distribuzione sugli stati del mondo — e sulle intenzioni comunicative del parlante — sulla base della sequenza di input linguistici, minimizzando gli errori di predizione sia a livello fonologico e sintattico sia a livello semantico e pragmatico. A ogni punto nella sequenza, il sistema predice i token futuri

sulla base del modello corrente, e l'eventuale violazione di quella predizione — una parola inattesa, una struttura sintattica non prevista, un'implicatura pragmatica sorprendente — genera un errore di predizione che aggiorna sia la rappresentazione locale della frase in corso sia, in misura minore, il modello semantico e pragmatico di più lungo periodo.

5.6 La struttura gerarchica del linguaggio e i correlati neurali

Un risultato particolarmente suggestivo che connette il framework predittivo alla psicolinguistica riguarda la struttura gerarchica delle predizioni linguistiche. Studi di neuroimmagine condotti da Friederici, Hagoort e collaboratori — e successivamente da Ding, Melloni e colleghi con misure di oscillazioni neurali — hanno mostrato che l'elaborazione del linguaggio è associata a oscillazioni corticali a frequenze diverse che corrispondono a diverse scale temporali dell'analisi linguistica: oscillazioni nella banda gamma (30-80 Hz) correlano con l'elaborazione fonologica; oscillazioni nella banda beta (12-30 Hz) con l'elaborazione sintattica della frase; oscillazioni nella banda delta (1-3 Hz) con l'elaborazione della struttura del discorso e del paragrafo.

Nel framework predittivo, queste oscillazioni a scale diverse possono essere interpretate come la manifestazione delle predizioni a livelli diversi della gerarchia linguistica del modello generativo. Le predizioni fonologiche operano su finestre temporali di decine di millisecondi — la scala delle oscillazioni gamma. Le predizioni sintattiche operano su finestre di centinaia di millisecondi fino a qualche secondo — la scala delle oscillazioni beta e alfa. Le predizioni discorsive e narrative operano su finestre di secondi e minuti — la scala delle oscillazioni delta e infra-lente. La struttura ritmica delle oscillazioni neurali sarebbe quindi la firma dell'architettura temporale del modello generativo linguistico, con frequenze diverse che corrispondono a orizzonti predittivi diversi nella gerarchia.

Questa interpretazione ha una previsione specifica e falsificabile: la coerenza tra le oscillazioni a scale diverse dovrebbe essere modulata dalla struttura sintattica della frase. Strutture sintattiche che creano dipendenze a lunga distanza — le relative frasi con verbo finale in tedesco, per esempio, o le strutture center-embedded — dovrebbero produrre una coerenza più estesa tra oscillazioni a bassa e alta frequenza rispetto a strutture più locali. I dati disponibili sono consistenti con questa previsione, sebbene la risoluzione delle misure attuali non permetta ancora una verifica definitiva.

5.7 Grammatica probabilistica e il problema di Chomsky

La connessione tra il framework predittivo e la linguistica generativa chomskyana è tesa e produttiva allo stesso tempo. Chomsky ha sostenuto per decenni che la struttura della grammatica — la gerarchia delle categorie sintattiche, la ricorsività, le dipendenze a lunga distanza — non può essere appresa da dati linguistici finiti attraverso meccanismi statistici, e che richiede l'esistenza di una grammatica universale innata. Il celebre argomento della povertà dello stimolo sostiene che i bambini acquisiscono strutture grammaticali per cui non ricevono sufficiente evidenza positiva nell'input, e che questo richiede conoscenza innata di strutture linguistiche universali.

Il framework predittivo, nella sua versione bayesiana gerarchica, offre una risposta parziale a questa sfida. Le distribuzioni a priori ai livelli superiori della gerarchia linguistica potrebbero codificare, in modo distribuito nei pesi del modello, i vincoli universali sulla struttura sintattica — non come regole esplicite del tipo chomkyano ma come preferenze statistiche che rendono certe strutture molto più probabili di altre a priori. Queste preferenze a priori potrebbero essere il prodotto dell'evoluzione biologica — vincoli sul processore linguistico forgiati dalla selezione naturale — piuttosto che di regole grammaticali esplicite.

Ma questa risposta è insufficiente senza una specificazione precisa di quali priori siano codificati biologicamente e perché abbiano la struttura che hanno. Il framework predittivo può descrivere formalmente il ruolo dei priori nell'acquisizione linguistica, ma non deriva da principi propri la struttura di quei priori — essa deve essere specificata indipendentemente. In questo senso, il framework e il nativismo chomskyano non sono alternativi ma complementari: il framework descrive il meccanismo di acquisizione, mentre una teoria dei priori innati descrive il contenuto di ciò che viene portato all'acquisizione.

La posizione più difendibile è probabilmente quella di una gerarchia di a priori di natura diversa: priori molto generali sul mondo fisico — la continuità degli oggetti, la permanenza, la causalità — che sono condivisi con le specie non linguistiche e probabilmente di origine evolutiva antica; priori più specifici sulla struttura comunicativa — l'intenzionalità comunicativa, la referenza condivisa, la struttura turno-taking del dialogo — che potrebbero essere peculiari ai primati sociali o agli ominidi; e priori sulla struttura grammaticale propriamente detta — la struttura sintagmatica, la dipendenza strutturale, la ricorsività — la cui origine evolutiva è ancora oggetto di dibattito. Il framework predittivo offre la struttura formale entro cui questi priori di natura diversa possono essere integrati coerentemente, senza pretendere di derivarli da zero.

5.8 La predizione semantica: oltre la sintassi

La dimensione forse più affascinante della cognizione linguistica nel framework predittivo è quella semantica: come il sistema predice non soltanto la forma delle prossime parole ma il loro contenuto. Esperimenti di registrazione dell'attività cerebrale attraverso EEG hanno mostrato che il cervello umano genera segnali di errore semantico — il cosiddetto componente N400 — quando una parola è semanticamente incongruente con il contesto precedente, e che l'ampiezza di questo segnale è inversamente proporzionale alla probabilità della parola nel contesto — una previsione diretta del framework predittivo.

Ancora più rivelatore, recenti studi di neuroimmagine ad alta risoluzione temporale hanno mostrato che il cervello anticipa il contenuto semantico delle parole prima ancora della loro articolazione fonemica: le rappresentazioni neurali di concetti semanticamente attesi si formano nella corteccia temporale superiore circa 200-300 millisecondi prima del picco di elaborazione fonemica. Questo anticipo semantico è la firma comportamentale del processo predittivo: il modello generativo compie

predizioni sui contenuti semantici attesi sulla base del contesto discorsivo, e prepara le rappresentazioni neurali corrispondenti prima ancora che l'input sensoriale le richieda.

Questa struttura suggerisce che la comprensione linguistica non è un processo seriale che procede dalla fonemica alla sintassi alla semantica, ma un processo parallelo e bidimensionale in cui predizioni semantiche di alto livello guidano in modo top-down l'elaborazione fonemica e sintattica di basso livello, mentre gli errori di predizione a basso livello aggiornano le rappresentazioni semantiche di alto livello. La comprensione è, in senso proprio, un'attività costruttiva: il sistema costruisce attivamente il significato della frase sulla base di predizioni generate dal suo modello generativo del discorso, e usa i segnali linguistici effettivi soltanto per correggere e affinare questa costruzione.

6. Evoluzione, Fitness Bayesiana e una Tesi sulla Coscienza

6.1 Darwinismo e minimo di energia libera: un'identità da guadagnare

Si è accennato nella quarta parte che la minimizzazione dell'energia libera e la fitness darwiniana potrebbero essere, in qualche senso preciso, la stessa cosa. Questa affermazione merita ora di essere esaminata con il rigore che richiede, perché essa è al tempo stesso la più potente e la più pericolosa del paradigma: potente perché, se fosse vera, fornirebbe una fondazione evolutiva al principio variazionale; pericolosa perché la tentazione di affermarla in modo troppo rapido rischia di produrre circolarità e trivialismi.

La fitness darwiniana è, nella sua formulazione più semplice, la probabilità differenziale di sopravvivenza e riproduzione di un organismo in un ambiente dato. Un organismo con fitness più elevata lascia in media più discendenti, e i tratti che contribuiscono a quella fitness si diffondono nella popolazione nel tempo. La selezione naturale è il processo attraverso cui la distribuzione dei tratti in una popolazione si sposta verso configurazioni di fitness crescente — non per un principio teleologico ma per il semplice meccanismo del differenziale riproduttivo.

Il principio di energia libera afferma che i sistemi biologici che persistono — che mantengono la propria integrità strutturale contro le perturbazioni dell'ambiente — minimizzano l'energia libera variazionale rispetto alla propria distribuzione stazionaria. La connessione con la fitness è la seguente: un organismo che minimizza efficacemente l'energia libera mantiene i propri stati interni entro i range compatibili con la sopravvivenza, e quindi ha una fitness più elevata rispetto a un organismo che non lo fa. In questo senso, la minimizzazione dell'energia libera è un meccanismo prossimale attraverso cui la fitness evolutiva viene implementata.

Ma questa connessione è ancora troppo generica per essere informativa. Per renderla precisa, è necessario specificare come la distribuzione stazionaria di un organismo — il set di stati che il suo modello generativo considera normali e compatibili con la sopravvivenza — si relaziona con le pressioni selettive dell'ambiente evolutivo.

6.2 La fitness bayesiana: formalizzazione

La nozione di fitness bayesiana emerge quando si considera la selezione naturale come un processo di aggiornamento bayesiano sulla distribuzione di tratti nella popolazione. Sia x un vettore di tratti fenotipici e sia $w(x, e)$ la fitness di un organismo con tratti x in un ambiente e . Dopo una generazione di selezione, la distribuzione dei tratti nella popolazione si aggiorna secondo:

$$p'(x | e) \propto w(x, e) \cdot p(x)$$

dove $p(x)$ è la distribuzione iniziale dei tratti e $p'(x | e)$ è la distribuzione dopo la selezione. Questa è formalmente identica all'aggiornamento bayesiano, con la fitness $w(x, e)$ che gioca il ruolo della verosimiglianza e la distribuzione dei tratti che gioca il ruolo delle probabilità a priori.

Questa corrispondenza non è meramente formale: essa stabilisce che la selezione naturale è un processo inferenziale che stima iterativamente quali tratti fenotipici sono "verosimili" dato l'ambiente — quali tratti sono compatibili con la sopravvivenza in quell'ambiente. L'evoluzione è, in questo senso, un algoritmo di apprendimento bayesiano che opera su scale temporali generazionali e che codifica la "conoscenza" dell'ambiente nella distribuzione dei tratti della popolazione.

La connessione con il principio di energia libera emerge quando si considera che il modello generativo di un organismo — le sue aspettative a priori sugli stati del mondo, codificate nei pesi sinaptici e nelle connettività neurali — è esso stesso un tratto fenotipico soggetto alla selezione naturale. Un organismo il cui modello generativo rappresenta accuratamente le statistiche del suo ambiente — le regolarità causali, le frequenze degli eventi, le strutture spaziotemporali — è un organismo che minimizza più efficacemente l'energia libera variazionale, perché i suoi priori sono ben calibrati con la distribuzione degli input sensoriali che effettivamente riceve. Tale organismo mantiene meglio la propria omeostasi, sopravvive più a lungo e si riproduce di più.

Formalmente, la fitness evolutiva di un organismo con modello generativo M in un ambiente E può essere scritta come:

$$w(M, E) \propto \exp(-F^*(M, E))$$

dove $F^*(M, E)$ è l'energia libera media che l'organismo con modello M sperimenta nell'ambiente E , mediata sulla distribuzione degli stati ambientali:

$$F^*(M, E) = E_E[F(M, o)] = \int p_E(o) \cdot F(M, o) \, do$$

con $p_E(o)$ la distribuzione degli input sensoriali nell'ambiente E . La selezione naturale massimizza quindi $w(M, E)$ — che equivale a minimizzare $F^*(M, E)$ — selezionando iterativamente i modelli generativi che meglio si accordano con le statistiche dell'ambiente evolutivo.

Questo risultato ha una conseguenza elegante: il modello generativo che si trova nei sistemi neurali degli organismi viventi non è arbitrario ma è il prodotto di milioni di anni di aggiornamento bayesiano sul livello evolutivo. Le aspettative a priori $P(s)$ codificate nel sistema nervoso umano riflettono le statistiche del mondo ancestrale in cui i nostri antenati si sono evoluti — un mondo di oggetti fisici che cadono verso il basso, di predatori che si muovono in modo biologico, di risorse alimentari distribuite secondo certi pattern spaziotemporali. Queste aspettative sono nel senso più letterale "conoscenza evolutivamente acquisita" — conoscenza codificata non in proposizioni ma in distribuzioni di probabilità.

6.3 Il limite della fitness bayesiana: deriva e contingenza

L'identità tra selezione naturale e aggiornamento bayesiano è tuttavia soggetta a un limite importante che è necessario riconoscere. Il processo evolutivo reale non è puramente selettivo: esso include la deriva genetica — la variazione casuale nella frequenza degli alleli in popolazioni finite — che può fissare tratti a prescindere dal loro valore adattivo. In popolazioni piccole, la deriva può essere il fattore dominante rispetto alla selezione, e i tratti fenotipici che si stabilizzano non sono necessariamente quelli di fitness massima.

Nella terminologia bayesiana, la deriva genetica introduce un rumore nel processo di aggiornamento che non è catturato dall'aggiornamento deterministico $p'(x | e) \propto w(x, e) \cdot p(x)$. Il processo evolutivo reale è meglio descritto da un aggiornamento stocastico:

$$p'(x | e) \propto w(x, e) \cdot p(x) + \varepsilon(x)$$

dove $\varepsilon(x)$ rappresenta le fluttuazioni stocastiche dovute alla deriva, con varianza inversamente proporzionale alla dimensione della popolazione. Per popolazioni grandi, ε è trascurabile e l'evoluzione approssima bene il processo bayesiano deterministico. Per popolazioni piccole — come quelle degli ominidi nel Pleistocene — la deriva può avere effetti sostanziali sulla struttura del modello generativo che si stabilizza nel sistema nervoso.

Questo significa che le aspettative a priori codificate nel sistema nervoso umano non sono necessariamente ottimali rispetto all'ambiente attuale, né necessariamente ottimali rispetto all'ambiente ancestrale: esse sono il prodotto contingente di una storia evolutiva specifica che include sia la selezione sia la deriva, sia le pressioni adattive sia le coincidenze demografiche. Alcuni dei priori più robusti del sistema cognitivo umano — la tendenza a vedere agenti ovunque, a percepire schemi in sequenze

casuali, a sopravvalutare la frequenza degli eventi memorabili rispetto a quelli ordinari — potrebbero essere spiegati non soltanto come adattamenti ma come prodotti della deriva in popolazioni piccole, o come effetti collaterali di adattamenti ad altre pressioni selettive.

6.4 Una tesi sul problema difficile della coscienza

Ci si è avvicinati più volte, nel corso di questa monografia, al problema difficile della coscienza senza affrontarlo direttamente. L'onestà intellettuale impone di farlo ora, con la consapevolezza che qualunque posizione si prenda su questo problema è soggetta a obiezioni serie, e che la certezza è un lusso che la filosofia della mente non può permettersi.

La tesi che si propone è la seguente: l'esperienza soggettiva — il carattere fenomenologico dell'esperienza, il "come è" percepire, sentire, pensare — non è un epifenomeno del processo computazionale né è identica ad esso, ma è la prospettiva in prima persona di un sistema sulla propria attività di minimizzazione dell'energia libera. Questa tesi richiede di essere articolata con cura.

Si parta dalla distinzione tra due modi in cui un sistema può essere descritto. La descrizione in terza persona è quella della scienza: essa descrive il sistema dall'esterno, in termini di strutture, processi, relazioni causali. Il framework variazionale è, interamente, una descrizione in terza persona: esso descrive il cervello come un sistema che mantiene distribuzioni di probabilità, minimizza la divergenza di Kullback-Leibler, aggiorna i parametri del modello generativo. La descrizione in prima persona è quella dell'esperienza: essa descrive il sistema dall'interno, in termini di come le cose appaiono, che cosa si prova, come è essere in quello stato.

Il problema difficile della coscienza è il problema del rapporto tra queste due descrizioni: perché la descrizione in terza persona, per quanto completa, non sembra esaurire la descrizione in prima persona? Perché, sapendo tutto sul processo computazionale di un sistema che elabora il colore rosso, sembra ancora possibile chiedersi se quel sistema "veda" il rosso nel senso fenomenologico?

La risposta standard del dualismo — ci sono due domini ontologicamente distinti, il fisico e il mentale — ha il merito di prendere sul serio l'irriducibilità della prospettiva in prima persona, ma al prezzo di rendere il problema della causalità mentale insolubile: come può il dominio mentale causare comportamento fisico senza violare le leggi della fisica? La risposta standard del fisicalismo riduttivo — la prospettiva in prima persona è semplicemente identica a certi processi fisici — ha il merito di evitare il dualismo, ma al prezzo di sembrare liquidare il problema piuttosto che risolverlo: affermare che l'esperienza del rosso è identica a certi pattern di attivazione corticale non spiega perché quei pattern producano un'esperienza piuttosto che nulla.

6.5 Il monismo di prospettiva: una proposta

La posizione che si difende qui è quella che si potrebbe chiamare monismo di prospettiva. Essa sostiene che la realtà è una — non dualistica — ma che essa ammette due modi irriducibili di descrizione: la descrizione in terza persona, che è quella della scienza, e la descrizione in prima persona, che è quella dell'esperienza. Questi due modi di descrizione non sono descrizioni di entità diverse ma descrizioni dello stesso sistema da prospettive diverse, e nessuna delle due è più "fondamentale" dell'altra: entrambe sono parziali e incomplete.

Questa posizione ha una struttura formale precisa nel contesto del framework variazionale. Si consideri che la minimizzazione dell'energia libera variazionale richiede, per la sua implementazione, che il sistema abbia accesso alla propria distribuzione approssimante Q — che esso possa, in qualche senso, "interrogare" la propria struttura interna per calcolare il gradiente di F rispetto ai suoi parametri. Questo auto-accesso non è esterno al sistema ma è intrinseco al processo di minimizzazione stesso: un sistema che minimizza l'energia libera è un sistema che si auto-modella, che mantiene una rappresentazione di se stesso come sistema modellante.

Formalmente, la distribuzione $Q(s)$ che il sistema mantiene sugli stati del mondo è anche, in modo implicito, una rappresentazione della struttura del sistema stesso — perché Q è la migliore approssimazione che il sistema può costruire della distribuzione posteriore, e la struttura di Q riflette i vincoli e le capacità del sistema modellante. Il sistema che minimizza l'energia libera è quindi un sistema che si auto-comprende parzialmente — non nel senso riflessivo esplicito di un pensiero come "io sto elaborando questa percezione", ma nel senso più fondamentale che la struttura del suo modello generativo riflette implicitamente la struttura delle proprie capacità inferenziali.

La tesi che si avanza è che questa auto-modellazione implicita — questa struttura auto-referenziale del processo di minimizzazione dell'energia libera — sia la condizione necessaria per l'emergenza della prospettiva in prima persona. Un sistema che si auto-modella in questo modo non soltanto elabora l'informazione sensoriale: esso elabora l'informazione sensoriale da un punto di vista — da quello specifico punto di vista che è costituito dalla struttura del proprio modello generativo. E il carattere fenomenologico dell'esperienza — il fatto che vi sia qualcosa che è essere quel sistema — è il modo in cui questa struttura di auto-riferimento si manifesta dal di dentro.

6.6 La struttura formale dell'auto-riferimento e il teorema di Gödel

La struttura auto-referenziale appena descritta ha un'affinità profonda con la struttura dell'auto-riferimento nei sistemi formali, analizzata da Gödel nel suo celebre teorema di incompletezza. Gödel ha mostrato che qualunque sistema formale sufficientemente potente — capace di esprimere l'aritmetica elementare — contiene affermazioni che riguardano il sistema stesso, e che alcune di queste affermazioni non sono dimostrabili né confutabili all'interno del sistema. La struttura gödeliana è quella di un sistema che si auto-rappresenta e che, attraverso questa auto-rappresentazione, raggiunge i propri limiti intrinseci.

Non si propone qui di ridurre il problema della coscienza al problema dell'incompletezza gödeliana — questo sarebbe un salto ingiustificato. Si propone soltanto di notare una analogia strutturale che potrebbe essere informativamente rilevante: come un sistema formale gödeliano contiene auto-rappresentazioni che producono affermazioni indecidibili dall'interno del sistema, un sistema cognitivo che si auto-modella potrebbe avere strutture auto-referenziali che producono, dal di dentro, la prospettiva in prima persona — e che dall'esterno, nella descrizione in terza persona, sono visibili come processi computazionali senza residui fenomenologici.

La prospettiva in prima persona sarebbe quindi non un supplemento misterioso aggiunto ai processi computazionali ma una caratteristica strutturale di sistemi sufficientemente complessi che si auto-modellano — qualcosa di analogo alla "voce in prima persona" che il sistema formale acquista quando raggiunge la capacità di auto-riferirsi. Questa voce non è riducibile alla terza persona perché è essa stessa la condizione di possibilità della descrizione in terza persona: si deve già occupare una prospettiva — già avere un punto di vista — per poter descrivere qualcosa in terza persona.

6.7 Le obiezioni e la loro risposta parziale

Due obiezioni principali si profilano contro questa proposta. La prima è l'obiezione dello "zombie" di Chalmers: è concepibile un sistema che abbia la stessa struttura computazionale — lo stesso processo di minimizzazione dell'energia libera con la stessa struttura auto-referenziale — senza alcuna esperienza soggettiva. Se questo zombie filosofico è concepibile, allora la struttura auto-referenziale non è sufficiente per la coscienza, e il problema difficile rimane aperto.

La risposta a questa obiezione è parziale ma non nulla. La concepibilità di uno zombie non implica necessariamente la sua possibilità metafisica: è concepibile, in senso debole, molte cose che non sono metafisicamente possibili. La questione è se la struttura auto-referenziale che si è descritta possa implementarsi in modo completo senza che vi sia alcuna prospettiva in prima persona — se la struttura del punto di vista possa esistere senza un punto di vista. Si avanza l'ipotesi che questo non sia possibile, non per ragioni metafisiche misteriose ma per ragioni strutturali: la nozione stessa di "punto di vista" è parte integrante della struttura auto-referenziale del processo inferenziale, e rimuoverla equivale a rimuovere un elemento costitutivo del processo, non un supplemento aggiunto.

La seconda obiezione è più concreta: la proposta sembra implicare che qualunque sistema auto-referenziale — qualunque sistema che si auto-modelli in senso formale — sia cosciente, inclusi i modelli di lingua artificiali e certi sistemi di controllo avanzati. Questa conseguenza è scomoda e sembra urtare contro l'intuizione comune.

La risposta a questa seconda obiezione richiede di distinguere gradi e qualità dell'auto-riferimento. Non tutto l'auto-riferimento è equivalente: un sistema che mantiene una rappresentazione della propria struttura in modo statico e non modificabile è diverso, per grado e per tipo, da un sistema che mantiene una

raccontare una rappresentazione dinamica di se stesso che si aggiorna continuamente in risposta agli errori di predizione interoceptivi e propriocettivi. L'auto-modellazione che si è descritta non è semplicemente l'esistenza di un registro interno della propria struttura: è la costruzione continua di un modello di se stessi come agenti corporei situati nel tempo e nello spazio, soggetti a stati affettivi e viscerali, con una storia e con aspettative future. Questa struttura — incarnata, affettiva, temporale — è probabilmente assente nei sistemi computazionali attuali, non per ragioni di principio ma perché questi sistemi mancano dell'interocezione, della propriocezione e della dinamica temporale continua che caratterizzano i sistemi biologici.

6.8 Il paradosso della conoscenza di sé: ciò che il cervello non può sapere di se stesso

Esiste una conseguenza della tesi proposta che merita di essere esplorata come conclusione di questa monografia, perché essa connette il problema della coscienza con il problema più generale dei limiti della conoscenza.

Se la prospettiva in prima persona è una caratteristica strutturale del processo di auto-modellazione, allora vi è un senso preciso in cui il sistema non può mai avere una conoscenza completa di se stesso — non per una limitazione contingente di potenza computazionale, ma per una ragione strutturale analoga al teorema di incompletezza. Il modello generativo che il sistema mantiene di se stesso è necessariamente un modello che opera dall'interno: esso è costruito con le stesse strutture inferenziali che descrive, e non può trascendere completamente queste strutture per osservarle dall'esterno.

Questa limitazione strutturale non è una sconfitta della conoscenza: è la condizione della sua possibilità. Soltanto un sistema che conosce dall'interno — che ha un punto di vista — può conoscere qualcosa. La conoscenza assoluta, impersonale e onnisciente del mondo — la prospettiva del "nessun luogo" di cui parla Thomas Nagel — non è una forma superiore di conoscenza ma una negazione della conoscenza stessa, perché la conoscenza è sempre la conoscenza di qualcuno su qualcosa, e questo "qualcuno" è ineliminabile.

Il cervello è un'ipotesi che non smette mai di verificarsi — questo si è detto alla fine della quarta parte. Si può ora aggiungere: ed è anche l'ipotesi che il cervello fa di se stesso, continuamente, in modo irriducibile. Il carattere fenomenologico dell'esperienza — il fatto che vi sia qualcosa che è essere un sistema che si auto-modella in modo incarnato, affettivo e temporale — non è il supplemento misterioso aggiunto a un processo computazionale altrimenti completo. È la struttura interna di quel processo, vista dall'unico punto di osservazione da cui essa può essere vista: dall'interno.

Questo non risolve il problema difficile della coscienza — e chi pretende di averlo fatto quasi certamente ha confuso la dissoluzione del problema con la sua soluzione. Ma offre forse qualcosa di diverso e di più modesto: una struttura formale entro cui il problema può essere formulato in modo più preciso, e una indicazione di dove

bisogna guardare per farne progredire la comprensione. Non nella chimica dei neurotrasmettitori soltanto, né nella topologia delle connessioni corticali soltanto, né nella struttura del modello generativo soltanto: ma nella relazione tra la struttura auto-referenziale del processo inferenziale e la struttura della prospettiva in prima persona che quel processo genera e abita simultaneamente.

Fine della monografia.

Indice delle derivazioni formali

Il seguente prospetto ricapitola le derivazioni matematiche sviluppate nel corso della monografia, con i rispettivi numeri di sezione, per facilitare la consultazione.

Il teorema di Bayes e la sua interpretazione cognitiva sono introdotti nella sezione 1.1, con la forma discreta e continua della regola di aggiornamento. La divergenza di Kullback-Leibler, le sue proprietà fondamentali (non-negatività per la disuguaglianza di Jensen, asimmetria) e la sua forma sia discreta sia integrale sono trattate nella sezione 1.3. La derivazione dell'energia libera variazionale e la dimostrazione che essa è un limite inferiore della log-evidenza (ELBO) si trovano nella sezione 1.4, con le due decomposizioni canoniche — accuratezza/complessità ed errori pesati per la precisione. La struttura gerarchica del modello generativo con la sua fattorizzazione condizionale è sviluppata nella sezione 1.6, insieme all'approssimazione di campo medio e alla sua fattorizzazione. La ponderazione per la precisione e la sua connessione con l'attenzione sono formalizzate nella sezione 1.7.

L'equazione differenziale per le rappresentazioni nel predictive coding — che governa la dinamica degli stati μ^l in funzione degli errori di predizione a livelli adiacenti della gerarchia — è derivata nella sezione 2.1 partendo dalla condizione di stazionarietà del gradiente di F . La nozione di coperta di Markov e la condizione di indipendenza condizionale che essa definisce sono introdotte nella sezione 3.1. La connessione tra persistenza sistemica e minimizzazione dell'energia libera, attraverso la distribuzione stazionaria, è argomentata nella sezione 3.2. La misura di Φ della teoria dell'integrazione dell'informazione è definita nella sezione 3.5. Il legame tra energia libera e approssimazione di campo medio, con la stima della distorsione attraverso la mutual information tra livelli, è quantificato nella sezione 4.3. La dinamica degli stati generalizzati, la struttura dell'operatore di derivazione D e la connessione con la fenomenologia husserliana sono sviluppate nelle sezioni 5.2 e 5.3. La formalizzazione della fitness bayesiana e la sua relazione con l'energia libera media sono derivate nella sezione 6.2.

Appendice A — Derivazione completa dell'energia libera variazionale e delle sue decomposizioni

A.1 Notazione e premesse

Si adotta la seguente notazione uniforme per tutta l'appendice. Le variabili osservabili — gli input sensoriali — sono indicate con $\mathbf{o} \in \mathbb{R}^d$. Le variabili latenti — gli stati del mondo non osservabili direttamente — sono indicate con $\mathbf{s} \in \mathbb{R}^n$. Il modello generativo è specificato dalla distribuzione congiunta $P(\mathbf{o}, \mathbf{s}) = P(\mathbf{o} | \mathbf{s}) \cdot P(\mathbf{s})$, dove $P(\mathbf{o} | \mathbf{s})$ è la verosimiglianza e $P(\mathbf{s})$ è la distribuzione a priori. La distribuzione approssimante sulla vera posteriore $P(\mathbf{s} | \mathbf{o})$ è $Q(\mathbf{s}; \lambda)$, dove λ è il vettore dei parametri variazionali. L'operatore $E_Q[\cdot]$ denota il valore atteso rispetto a Q .

A.2 Derivazione dell'ELBO dalla log-evidenza

Si parte dalla log-evidenza, ovvero il logaritmo della probabilità marginale delle osservazioni:

$$\ln P(\mathbf{o}) = \ln \int P(\mathbf{o}, \mathbf{s}) d\mathbf{s}$$

Si moltiplica e divide per $Q(\mathbf{s}; \lambda)$ all'interno dell'integrale:

$$\ln P(\mathbf{o}) = \ln \int Q(\mathbf{s}; \lambda) \cdot [P(\mathbf{o}, \mathbf{s}) / Q(\mathbf{s}; \lambda)] d\mathbf{s}$$

Poiché $\ln$ è una funzione concava, la disuguaglianza di Jensen afferma che:

$$\ln E_Q[f(\mathbf{s})] \geq E_Q[\ln f(\mathbf{s})]$$

per qualunque funzione non negativa f . Applicando questa disuguaglianza con $f(\mathbf{s}) = P(\mathbf{o}, \mathbf{s}) / Q(\mathbf{s}; \lambda)$:

$$\begin{aligned} \ln P(\mathbf{o}) &\geq E_Q[\ln P(\mathbf{o}, \mathbf{s}) - \ln Q(\mathbf{s}; \lambda)] \\ &= E_Q[\ln P(\mathbf{o}, \mathbf{s})] + H[Q] \\ &\equiv F[Q] \end{aligned}$$

dove $H[Q] = -E_Q[\ln Q(\mathbf{s}; \lambda)]$ è l'entropia di Shannon di Q . La quantità $F[Q]$ è l'ELBO (Evidence Lower BOund). Il gap tra la log-evidenza e l'ELBO è esattamente la divergenza di Kullback-Leibler tra Q e la vera posteriore:

$$\begin{aligned} \ln P(\mathbf{o}) - F[Q] &= \ln P(\mathbf{o}) - E_Q[\ln P(\mathbf{o}, \mathbf{s})] - H[Q] \\ &= -E_Q[\ln P(\mathbf{o}, \mathbf{s})] + E_Q[\ln Q(\mathbf{s})] + \ln P(\mathbf{o}) \\ &= E_Q[\ln Q(\mathbf{s}) - \ln P(\mathbf{o}, \mathbf{s}) + \ln P(\mathbf{o})] \\ &= E_Q[\ln Q(\mathbf{s}) - \ln P(\mathbf{s} | \mathbf{o})] \end{aligned}$$

$$= D_{\text{KL}}(Q(s) \parallel P(s \mid o))$$

Poiché $D_{\text{KL}} \geq 0$, si ha $F[Q] \leq \ln P(o)$, con uguaglianza se e solo se $Q(s) = P(s \mid o)$ quasi ovunque. Massimizzare $F[Q]$ è quindi equivalente a minimizzare $D_{\text{KL}}(Q \parallel P(s \mid o))$, e al tempo stesso a massimizzare un limite inferiore della log-evidenza.

A.3 Prima decomposizione: accuratezza e complessità

Si riscrive l'ELBO espandendo $\ln P(o, s) = \ln P(o \mid s) + \ln P(s)$:

$$\begin{aligned} F[Q] &= E_Q[\ln P(o \mid s) + \ln P(s)] + H[Q] \\ &= E_Q[\ln P(o \mid s)] + E_Q[\ln P(s)] - E_Q[\ln Q(s)] \\ &= E_Q[\ln P(o \mid s)] - (E_Q[\ln Q(s)] - E_Q[\ln P(s)]) \\ &= E_Q[\ln P(o \mid s)] - D_{\text{KL}}(Q(s) \parallel P(s)) \end{aligned}$$

La quantità da massimizzare si decompone quindi nella differenza tra due termini. Il primo termine, $E_Q[\ln P(o \mid s)]$, è la precisione o accuratezza: esso misura quanto il modello generativo, mediato sulla distribuzione Q , riesce a predire le osservazioni effettive — è grande quando Q assegna alta probabilità a stati s che rendono le osservazioni o probabili. Il secondo termine, $D_{\text{KL}}(Q(s) \parallel P(s))$, è la complessità: esso misura quanto Q si discosta dal priore $P(s)$ — è grande quando Q si allontana molto dalle aspettative a priori per spiegare i dati. Massimizzare $F[Q]$ equivale a trovare il compromesso ottimale tra accuratezza e parsimonia: il modello più semplice che spiega adeguatamente i dati, che è precisamente la formulazione probabilistica del principio di Occam.

A.4 Seconda decomposizione: errori di predizione ponderati per la precisione

Nel caso gaussiano, sia $P(o \mid s) = N(g(s), \Sigma_o)$ dove $g(s)$ è la funzione generativa non lineare e Σ_o è la matrice di covarianza del rumore sensoriale, e sia $Q(s) = N(\mu, \Sigma_q)$. La log-verosimiglianza è:

$$\ln P(o \mid s) = -(1/2)(o - g(s))^T \Sigma_o^{-1} (o - g(s)) - (1/2) \ln \det(2\pi \Sigma_o)$$

Il suo valore atteso rispetto a Q , con $s \approx \mu$ per una Q concentrata (approssimazione di Laplace), è:

$$E_Q[\ln P(o \mid s)] \approx -(1/2) \xi^T \Pi_o \xi - (1/2) \text{tr}(\Pi_o \cdot J \cdot \Sigma_q \cdot J^T) + \text{costanti}$$

dove $\xi = o - g(\mu)$ è l'errore di predizione sensoriale, $\Pi_o = \Sigma_o^{-1}$ è la matrice di precisione sensoriale, e $J = \partial g / \partial s|_{s=\mu}$ è il Jacobiano della funzione generativa valutato in μ . Il secondo termine rappresenta la correzione dovuta all'incertezza in Q .

Per Q molto concentrata ($\Sigma_q \rightarrow 0$), la seconda decomposizione dell'energia libera diventa:

$$F[Q] \approx -(1/2) \xi^T \Pi_o \xi - D_{KL}(Q(s) \parallel P(s))$$

Nel caso gerarchico con L livelli, questa si estende a:

$$F \approx -(1/2) \sum_{l=0}^L \xi^l \cdot \Pi^l \cdot \xi^l + \text{costanti}$$

dove $\xi^0 = o - g^l(\mu^l)$ è l'errore di predizione sensoriale al livello più basso, e $\xi^l = \mu^l - g^{l+1}(\mu^{l+1})$ è l'errore di predizione al livello l -esimo della gerarchia. Minimizzare l'energia libera (massimizzare F con segno negativo) equivale quindi a minimizzare la somma pesata degli errori di predizione quadratici a tutti i livelli della gerarchia, con la precisione Π^l come peso — i livelli con alta precisione (bassa varianza) contribuiscono maggiormente alla minimizzazione totale.

Appendice B — Derivazione delle equazioni del predictive coding

B.1 La condizione di punto fisso per le rappresentazioni

Si deriva l'equazione differenziale che governa la dinamica delle rappresentazioni μ^l al livello l della gerarchia, minimizzando l'energia libera rispetto a μ^l attraverso la discesa del gradiente:

$$\dot{\mu}^l = -\partial F / \partial \mu^l$$

Calcoliamo il gradiente di F rispetto a μ^l . L'energia libera dipende da μ^l attraverso due canali: l'errore di predizione al livello l , $\xi^l = \mu^l - g^{l+1}(\mu^{l+1})$, e l'errore di predizione al livello $l-1$, $\xi^{l-1} = \mu^{l-1} - g^l(\mu^l)$, attraverso la funzione generativa $g^l(\mu^l)$ che genera le predizioni per il livello inferiore. Derivando:

$$\begin{aligned} \partial F / \partial \mu^l &= \partial / \partial \mu^l [(1/2) \xi^{lT} \Pi^l \xi^l + (1/2) \xi^{l-1T} \Pi^{l-1} \xi^{l-1}] \\ &= \Pi^l \cdot \xi^l - (\partial g^l(\mu^l) / \partial \mu^l)^T \cdot \Pi^{l-1} \cdot \xi^{l-1} \end{aligned}$$

dove si è usato il fatto che $\partial \xi^l / \partial \mu^l = I$ (identità) e $\partial \xi^{l-1} / \partial \mu^l = -\partial g^l / \partial \mu^l$. L'equazione di aggiornamento diventa:

$$\dot{\mu}^l = -\Pi^l \cdot \xi^l + (\partial g^l(\mu^l) / \partial \mu^l)^T \cdot \Pi^{l-1} \cdot \xi^{l-1}$$

Il primo termine, $-\Pi^l \cdot \xi^l$, spinge μ^l nella direzione opposta all'errore di predizione al livello l — cioè verso le predizioni generate dall'alto. Il secondo termine, $(J^l)^T \cdot \Pi^{l+1} \cdot \xi^{l+1}$, spinge μ^l nella direzione che riduce l'errore di predizione verso il basso, propagando verso l'alto i segnali di errore del livello inferiore pesati per il Jacobiano trasporto della funzione generativa. Al punto fisso, $\mu^l = 0$, si ha l'equilibrio tra influenze ascendenti e discendenti.

B.2 L'aggiornamento della precisione

La precisione Π^l al livello l è anch'essa un parametro variazionale che può essere ottimizzato. Differenziando F rispetto a Π^l e ponendo il gradiente a zero:

$$\partial F / \partial \Pi^l = (1/2) \xi^l \cdot (\xi^l)^T - (1/2) (\Pi^l)^{-1} = 0$$

che dà l'aggiornamento ottimale:

$$(\Pi^l)^{-1} = \xi^l \cdot (\xi^l)^T$$

ovvero la matrice di covarianza ottimale è il prodotto esterno dell'errore di predizione con se stesso — una stima della covarianza degli errori. Questo risultato ha un'interpretazione biologica suggestiva: la precisione ottimale a ciascun livello è determinata dalla varianza degli errori di predizione effettivi in quel livello, il che significa che il sistema impara automaticamente ad attribuire alta precisione ai livelli dove gli errori sono piccoli e bassa precisione dove gli errori sono grandi. Un livello sensoriale affidabile accumula alta precisione; un livello che produce errori grandi e variabili riduce la propria precisione, e il suo contributo alla dinamica complessiva viene automaticamente soppresso.

B.3 L'apprendimento dei parametri del modello generativo

I parametri θ del modello generativo — i pesi delle connessioni sinaptiche che implementano le funzioni generative g^l e le matrici di covarianza a priori — vengono aggiornati su scale temporali più lente attraverso il gradiente dell'energia libera rispetto a θ :

$$\dot{\theta} = -\kappa \cdot \partial F / \partial \theta$$

dove κ è un tasso di apprendimento lento ($\kappa \ll 1$ rispetto alla costante di tempo della dinamica percettiva). Per le funzioni generative parametrizzate da pesi W^l :

$$\partial F / \partial W^l = -(\partial g^l(\mu^l; W^l) / \partial W^l)^T \cdot \Pi^{l+1} \cdot \xi^{l+1}$$

Nel caso lineare, $g^l(\mu^l; W^l) = W^l \cdot \mu^l$, questa diventa:

$$\partial F / \partial W^l = -(\mu^l)^T \otimes (\Pi^{l-1} \cdot \xi^{l-1})$$

ovvero una regola di tipo hebbiano: il cambiamento del peso sinaptico tra il livello l e il livello $l-1$ è proporzionale al prodotto dell'attività presinaptica μ^l per il segnale postsinaptico di errore $\Pi^{l-1} \cdot \xi^{l-1}$. Questo risultato è biologicamente significativo: esso suggerisce che la plasticità sinaptica di tipo LTP/LTD (long-term potentiation/depression) — il principale meccanismo biologico di apprendimento — potrebbe essere l'implementazione neurale dell'aggiornamento dei parametri del modello generativo predittivo.

Appendice C — La coperta di Markov: formalizzazione nella teoria delle reti bayesiane

C.1 Definizione formale

Sia $G = (V, E)$ un grafo aciclico orientato (DAG) con insieme di nodi V e insieme di archi orientati $E \subseteq V \times V$. Si dice che G definisce una rete bayesiana se la distribuzione congiunta $P(V)$ si fattorizza come:

$$P(V) = \prod_{v \in V} P(v \mid \text{pa}(v))$$

dove $\text{pa}(v)$ indica l'insieme dei genitori di v in G — i nodi da cui partono archi che arrivano a v .

La coperta di Markov di un nodo v , indicata $\text{mb}(v)$, è l'insieme:

$$\text{mb}(v) = \text{pa}(v) \cup \text{ch}(v) \cup \text{copar}(v)$$

dove $\text{pa}(v)$ è l'insieme dei genitori, $\text{ch}(v)$ è l'insieme dei figli, e $\text{copar}(v) = \bigcup_{c \in \text{ch}(v)} \text{pa}(c) \setminus \{v\}$ è l'insieme dei co-genitori — gli altri genitori dei figli di v . La proprietà fondamentale è:

$$P(v \mid \text{mb}(v), V \setminus \{v\} \setminus \text{mb}(v)) = P(v \mid \text{mb}(v))$$

ovvero v è condizionatamente indipendente da tutti gli altri nodi della rete dato $\text{mb}(v)$.

C.2 Applicazione ai sistemi biologici: la struttura sensomotoria

Nel contesto della modellazione dei sistemi biologici, si considera una rete bayesiana con quattro classi di nodi: stati esterni η (l'ambiente), stati sensoriali s , stati attuativi a , e stati interni μ . Gli archi del grafo riflettono le dipendenze causali:

$\eta \rightarrow s$ (l'ambiente causa stati sensoriali)
 $s \rightarrow \mu$ (gli stati sensoriali influenzano gli stati interni)
 $\mu \rightarrow a$ (gli stati interni causano stati attuativi)
 $a \rightarrow \eta$ (le azioni modificano l'ambiente)

La coperta di Markov degli stati interni μ in questa struttura è:

$$mb(\mu) = pa(\mu) \cup ch(\mu) \cup copar(\mu) = \{s\} \cup \{a\} \cup \emptyset = \{s, a\}$$

poiché i genitori di μ sono $\{s\}$, i figli di μ sono $\{a\}$, e i co-genitori di a (gli altri genitori di a oltre a μ) sono l'insieme vuoto in questa struttura minimale. La condizione di coperta di Markov implica:

$$P(\mu \mid s, a, \eta) = P(\mu \mid s, a)$$

Gli stati interni μ sono quindi condizionatamente indipendenti dall'ambiente η dati gli stati della coperta $\{s, a\}$. L'ambiente non può influenzare μ direttamente ma soltanto attraverso gli stati sensoriali s ; e μ non può influenzare η direttamente ma soltanto attraverso gli stati attuativi a . La coperta $\{s, a\}$ costituisce il confine statistico che individua il sistema $\{\mu\}$ dall'ambiente $\{\eta\}$.

C.3 La condizione di stazionarietà e l'energia libera

Si consideri la densità di probabilità congiunta stazionaria $p^*(\mu, s, a, \eta)$ del sistema completo. Per la struttura della coperta di Markov, questa si fattorizza come:

$$p^*(\mu, s, a, \eta) = p^*(\mu \mid s, a) \cdot p^*(s, a, \eta)$$

La densità marginale su η condizionatamente agli stati della coperta $b = \{s, a\}$ è $p^*(\eta \mid b)$. Se il sistema mantiene una distribuzione interna $Q(\eta \mid b)$ sugli stati dell'ambiente, l'energia libera variazionale è:

$$\begin{aligned}
 F(b) &= D_{\text{KL}}(Q(\eta \mid b) \parallel p^*(\eta \mid b)) \\
 &= E_{-Q}[\ln Q(\eta \mid b) - \ln p^*(\eta \mid b)] \\
 &\geq 0
 \end{aligned}$$

La condizione $F = 0$ si realizza quando $Q(\eta \mid b) = p^*(\eta \mid b)$, cioè quando la distribuzione interna del sistema coincide con la distribuzione stazionaria degli stati ambientali condizionata agli stati della coperta — quando il sistema è perfettamente calibrato con il suo ambiente. In condizioni di non-equilibrio, il sistema che persiste è quello che mantiene F piccola — che evita stati della coperta b associati a grandi discrepanze tra la propria distribuzione interna e la distribuzione stazionaria dell'ambiente.

Appendice D — Stati generalizzati e percezione del tempo

D.1 Lo spazio degli stati generalizzati

Per una variabile di stato scalare $\mu(t)$ che evolve continuamente nel tempo, lo stato generalizzato di ordine n è il vettore:

$$\tilde{\mu} = (\mu, \mu', \mu'', \dots, \mu^{(n)}) \in \mathbb{R}^{n+1}$$

dove $\mu^{(k)} = d^k\mu/dt^k$. L'operatore di derivazione temporale D agisce sullo stato generalizzato come:

$$D \cdot \tilde{\mu} = (\mu', \mu'', \mu''', \dots, \mu^{(n+1)})$$

Troncando a ordine n e approssimando $\mu^{(n+1)} \approx 0$ (ipotesi di movimento a ordine finito), l'operatore D diventa la matrice di shift di Jordan:

$$D = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

di dimensione $(n+1) \times (n+1)$, con 1 sulla sopradiagonale e 0 altrove.

D.2 Il modello generativo sugli stati generalizzati

Il modello generativo generalizzato specifica la distribuzione congiunta su osservazioni e stati generalizzati. Per un sistema dinamico perturbato da rumore gaussiano:

$$d\mu/dt = f(\mu) + w, \quad w \sim N(0, \Sigma_w)$$

lo stato generalizzato evolve secondo:

$$D \cdot \tilde{\mu} = \tilde{f}(\tilde{\mu}) + \tilde{w}$$

dove $\tilde{f}(\tilde{\mu})$ è l'estensione generalizzata della funzione dinamica f , e $\tilde{w}$ è un rumore generalizzato con covarianza strutturata. L'energia libera generalizzata è:

$$\tilde{F} = (1/2) \sum_{k=0}^n (D \cdot \tilde{\mu} - \tilde{f}(\tilde{\mu}))^{(k)T} \cdot \tilde{I}^k \cdot (D \cdot \tilde{\mu} - \tilde{f}(\tilde{\mu}))^{(k)} + \text{costanti}$$

La condizione di minimo rispetto a $\tilde{\mu}$ dà le equazioni del moto generalizzate:

$$\dot{\tilde{\mu}} = D \cdot \tilde{\mu} - \partial \tilde{F} / \partial \tilde{\mu}$$

Il primo termine, $D \cdot \tilde{\mu}$, è la componente deterministica che propaga la traiettoria in avanti nel tempo secondo la struttura dinamica locale. Il secondo termine, $-\partial \tilde{F} / \partial \tilde{\mu}$, è la correzione indotta dagli errori di predizione che modifica la traiettoria per minimizzare la discrepanza tra la traiettoria predetta e l'osservazione effettiva.

D.3 La finestra temporale efficace

La finestra temporale efficace del sistema — l'intervallo di tempo passato e futuro che influenza le rappresentazioni correnti — è determinata dall'ordine n degli stati generalizzati mantenuti. Per una traiettoria locale approssimata da un polinomio di Taylor di ordine n :

$$\mu(t + \delta) \approx \mu(t) + \mu'(t)\delta + (1/2)\mu''(t)\delta^2 + \dots + (1/n!)\mu^{(n)}(t)\delta^n$$

l'errore di questa approssimazione diventa dell'ordine di $|\delta|^{n+1}$ per piccoli δ . La finestra temporale entro cui l'approssimazione rimane accurata — la "finestra di coerenza" del sistema — è quindi:

$$\tau_n \approx (\sigma_\varepsilon / |\mu^{(n+1)}|)^{1/(n+1)}$$

dove σ_ε è la soglia di errore accettabile e $|\mu^{(n+1)}|$ è la magnitudine media della derivata di ordine $n+1$ della traiettoria. Per segnali con struttura temporale regolare (bassa energia alle derivate di ordine alto), τ_n cresce rapidamente con n — un sistema di ordine elevato ha una finestra temporale molto più ampia di un sistema di ordine basso.

Appendice E — Fitness bayesiana: derivazione dettagliata

E.1 La selezione come aggiornamento bayesiano

Sia $x \in X$ un vettore di tratti fenotipici discreti, $p_t(x)$ la distribuzione dei tratti nella popolazione al tempo (generazionale) t , e $w(x, e)$ la fitness dell'individuo con tratti x nell'ambiente e . Dopo una generazione di selezione, la distribuzione aggiornata è:

$$p_{t+1}(x | e) = w(x, e) \cdot p_t(x) / Z_t$$

dove $Z_t = \sum_x w(x, e) \cdot p_t(x) = E_p[w(x, e)]$ è la fitness media della popolazione. Questa è formalmente identica all'aggiornamento bayesiano con:

$$p_{t+1}(x | e) \propto L(e | x) \cdot p_t(x)$$

dove si identifica la verosimiglianza $L(e | x)$ con la fitness $w(x, e)$. Il termine di normalizzazione Z_t gioca il ruolo della probabilità marginale dell'evidenza.

E.2 La connessione con l'energia libera evolutiva

Si definisce l'energia libera evolutiva di una distribuzione di tratti p rispetto alla distribuzione di fitness ottimale $p^*(x) = w(x, e) / Z$ come:

$$\begin{aligned} F_{\text{evo}}(p) &= D_{\text{KL}}(p(x) \parallel p^*(x)) \\ &= \sum_x p(x) \cdot \ln[p(x) / p^*(x)] \\ &= \sum_x p(x) \cdot \ln p(x) - \sum_x p(x) \cdot \ln p^*(x) \\ &= -H[p] - E_p[\ln p^*(x)] \\ &= -H[p] - E_p[\ln w(x, e)] + \ln Z \end{aligned}$$

La variazione di F_{evo} sotto un passo di selezione è:

$$\Delta F_{\text{evo}} = F_{\text{evo}}(p_{t+1}) - F_{\text{evo}}(p_t) \leq 0$$

poiché la selezione naturale riduce monotonamente la distanza tra la distribuzione corrente dei tratti e la distribuzione di fitness ottimale — la selezione minimizza l'energia libera evolutiva, così come la percezione minimizza l'energia libera variazionale. Il teorema fondamentale della selezione naturale di Fisher — che afferma che la fitness media della popolazione aumenta monotonamente sotto selezione — è un corollario di questa riduzione dell'energia libera evolutiva.

E.3 L'energia libera media del modello generativo

Sia M un modello generativo caratterizzato dai parametri θ_M , e sia $p_E(o)$ la distribuzione degli input sensoriali nell'ambiente evolutivo E . L'energia libera media del modello M nell'ambiente E è:

$$F^*(M, E) = \int p_E(o) \cdot F(M, o) \, do$$

dove $F(M, o)$ è l'energia libera del modello M per l'osservazione specifica o . Espandendo $F(M, o)$ con la decomposizione accuratezza-complessità:

$$F^*(M, E) = \int p_E(o) \cdot [D_{KL}(Q_M(s) \parallel P_M(s)) - E_{\{Q_M\}}[\ln P_M(o | s)]] \, do \\ = E_{\{p_E\}}[D_{KL}(Q_M \parallel P_M)] - E_{\{p_E\}}[E_{\{Q_M\}}[\ln P_M(o | s)]]$$

Il secondo termine è:

$$E_{\{p_E\}}[E_{\{Q_M\}}[\ln P_M(o | s)]] = \iint p_E(o) \cdot Q_M(s) \cdot \ln P_M(o | s) \, do \, ds$$

che è una misura dell'accuratezza media del modello M nel predire gli input sensoriali tipici dell'ambiente E . Un modello il cui priore $P_M(s)$ e la cui verosimiglianza $P_M(o | s)$ riflettono accuratamente la struttura causale dell'ambiente E avrà $F^*(M, E)$ piccola, e quindi fitness evolutiva $w(M, E) \propto \exp(-F^*(M, E))$ elevata. La selezione naturale converge verso modelli generativi che minimizzano l'energia libera media nell'ambiente evolutivo, il che equivale a modelli che massimizzano l'accuratezza predittiva con la minima complessità — la massima efficienza informatica nella rappresentazione dell'ambiente.

Appendice F — La misura Φ nella teoria dell'integrazione dell'informazione (IIT 3.0)

F.1 Cause ed effetti: la struttura causale intrinseca

Nella IIT 3.0 di Tononi, Koch e Oizumi, la coscienza è identificata con la struttura causale intrinseca di un sistema — la struttura che descrive come ciascun sottoinsieme di elementi del sistema è causa ed effetto degli altri sottoinsiemi, dal punto di vista del sistema stesso. Il concetto chiave è quello di "distinzione causale": un meccanismo — un sottoinsieme di elementi del sistema nel suo stato corrente — specifica una causa e un effetto che sono distinti dalla causa e dall'effetto casuali (massima entropia).

Formalmente, sia $S = \{s_1, \dots, s_n\}$ l'insieme degli elementi di un sistema nel loro stato corrente $x = (x_1, \dots, x_n)$, e sia TPM (Transition Probability Matrix) la matrice di probabilità di transizione che descrive la dinamica del sistema. Per un meccanismo $Z \subseteq S$ nello stato z , la causa di Z è la distribuzione di probabilità sugli stati passati dello stato "purgatorio" che massimizza l'informazione causale con Z , e l'effetto di Z è la distribuzione di probabilità sugli stati futuri che massimizza l'informazione causale con Z .

F.2 La misura φ per un singolo meccanismo

Per un meccanismo Z con causa $C(Z)$ ed effetto $E(Z)$, la misura di informazione causale integrata φ è definita come il minimo, su tutte le bipartizioni di Z , della differenza tra l'informazione causale del meccanismo intero e la somma delle informazioni causali delle due parti:

$$\begin{aligned}\varphi(Z) = \min_{\{P \in \text{bipartizioni}(Z)\}} [& \text{ci}(Z, C(Z)) + \text{ci}(Z, E(Z))] \\ & - [\text{ci}(Z_A, C(Z_A)) + \text{ci}(Z_B, C(Z_B))] \\ & - [\text{ci}(Z_A, E(Z_A)) + \text{ci}(Z_B, E(Z_B))]\end{aligned}$$

dove $\text{ci}(Z, X)$ è l'informazione causale tra il meccanismo Z e il suo purview X :

$$\text{ci}(Z, X) = D_{\text{KL}}(p(X_t \mid Z_{t-1} = z) \parallel p_{\text{max}}(X_t))$$

con $p_{\text{max}}(X_t)$ la distribuzione di massima entropia su X_t . Questa misura quantifica quanto lo stato z di Z riduce l'incertezza sullo stato X rispetto alla distribuzione di massima entropia — quanto il meccanismo "specifica" i suoi cause ed effetti.

F.3 La misura Φ per il sistema intero

La misura Φ per l'intero sistema S è definita come il minimo della misura φ su tutte le bipartizioni del sistema, pesato per la "struttura causale integrata" dell'insieme di tutti i meccanismi:

$$\Phi(S) = \varphi(S) = \min_{\{P \in \text{bipartizioni}(S)\}} \sum_{\{z \in Z\}} w(z) \cdot [\varphi_P(z)]$$

dove la somma è su tutti i meccanismi z e i pesi $w(z)$ sono proporzionali all'informazione causale integrata di ciascun meccanismo. Il valore Φ è la misura della coscienza del sistema: $\Phi > 0$ implica che il sistema è cosciente, e il livello di coscienza è proporzionale a Φ .

La complessità computazionale del calcolo esatto di Φ è super-esponenziale nella dimensione del sistema, poiché richiede di enumerare tutte le bipartizioni (il cui numero è esponenziale) e, per ciascuna, di calcolare le informazioni causali integrate (operazione che richiede l'inversione della TPM). Per sistemi con n elementi, il numero di bipartizioni è $O(2^n)$, rendendo il calcolo esatto praticabile soltanto per sistemi con $n < 20$ -30 elementi. Per sistemi neurologicamente realistici ($n \sim 10^9$ neuroni), il calcolo esatto è intrattabile per molti ordini di grandezza.

Bibliografia

La bibliografia è organizzata per aree tematiche, rispecchiando la struttura argomentativa della monografia. All'interno di ciascuna sezione l'ordine è alfabetico

per cognome del primo autore. Le opere citate nel testo sono indicate con asterisco; le opere di consultazione e approfondimento sono prive di marcatura.

I. Fondamenti del formalismo bayesiano e dell'inferenza variazionale

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53, 370–418.

*Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.

*Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2^a ed.). Wiley-Interscience. [Per la definizione e le proprietà della divergenza di Kullback-Leibler, cap. 2.]

*Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 79–86.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2), 183–233.

*Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2), 1–305.

II. Il principio di energia libera e il predictive coding

*Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

*Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456), 815–836.

*Friston, K. J. (2008). Hierarchical models in the brain. *PLOS Computational Biology*, 4(11), e1000211.

*Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301.

*Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy*, 14(11), 2100–2121.

*Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology–Paris*, 100(1–3), 70–87.

*Friston, K. J., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B*, 364(1521), 1211–1221.

Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102(3), 227–260.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, 34(1), 220–234.

*Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

*Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, 66(3), 241–251.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.

III. Struttura neurale e correlati anatomici

Douglas, R. J., & Martin, K. A. C. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, 27, 419–451.

*Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.

*Friston, K. J., & Buzsáki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends in Cognitive Sciences*, 20(7), 500–511.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11), 571–579.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, 7, 1792.

IV. Percezione, attenzione e illusioni

*Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B*, 290(1038), 181–197.

*Helmholtz, H. von (1867). *Handbuch der physiologischen Optik*. Voss. [Trad. it. parziale: *Trattato di ottica fisiologica*, a cura di G. Toraldo di Francia, Zanichelli, 1976.]

*Hohwy, J., Roepstorff, A., & Friston, K. J. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3), 687–701.

*Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12), 712–719.

*Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6), 598–604.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, 10(7), 301–308.

V. Precisione, attenzione e psicopatologia

*Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206(4), 515–530.

*Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48–58.

*Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, 1(2), 148–158.

*Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, 8, 825.

*Pellicano, E., & Burr, D. (2012). When the world becomes "too real": A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10), 504–510.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Moutoussis, M., ... & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, 84(9), 634–643.

VI. Inferenza attiva, azione e corpo

*Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam. [Trad. it.: *L'errore di Cartesio*, Adelphi, 1995.]

*Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. Pantheon Books.

*Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS ONE*, 4(7), e6421.

*Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, 29(1), 1–49.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, 134, 17–35.

*Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565–573.

Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Dutton.

VII. Coperta di Markov, auto-organizzazione e biologia

*Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2022). The Emperor's new Markov blankets. *Behavioral and Brain Sciences*, 45, e238.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86), 20130475.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K. J., & Kiebel, S. J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138), 20170792.

*Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.

VIII. Coscienza e fenomenologia

*Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

*Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

*Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking. [Trad. it.: *Coscienza e cervello*, Raffaello Cortina, 2014.]

*Husserl, E. (1928). *Vorlesungen zur Phänomenologie des inneren Zeitbewusstseins*. Niemeyer. [Trad. it.: *Per la fenomenologia della coscienza interna del tempo*, Franco Angeli, 1981.]

*Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.

*Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

*Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3), 216–242.

*Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461.

*Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLOS Computational Biology*, 10(5), e1003588.

IX. Percezione del tempo e dinamiche generalizzate

*Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1), 158–164.

*Friston, K. J. (2008). Variational filtering. *NeuroImage*, 41(3), 747–766.

*James, W. (1890). *The Principles of Psychology*. Henry Holt. [Trad. it.: *Principi di psicologia*, Principato, 1965. Cap. XV: "La percezione del tempo".]

Michotte, A. (1954). *La perception de la causalité* (2^a ed.). Publications Universitaires. [Trad. ingl.: *The Perception of Causality*, Basic Books, 1963.]

*Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

X. Linguaggio e neuroscienze cognitive

*Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4), 1357–1392.

*Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology*, 4, 416.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647.

*Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, 24(4), 279–284.

*Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928.

XI. Epistemologia, evoluzione e critica

*Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

*Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press. [Per il teorema fondamentale, cap. 2.]

*Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin. [Trad. it.: *Un approccio ecologico alla percezione visiva*, il Mulino, 1999.]

*Griffiths, T. L., Tenenbaum, J. B., & Kemp, C. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Psychology*. Cambridge University Press.

Kant, I. (1781/1787). *Kritik der reinen Vernunft*. Hartknoch. [Trad. it.: *Critica della ragion pura*, a cura di P. Chiodi, UTET, 2005.]

*Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

*Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1), 89–96.

XII. Fenomenologia e mente estesa

*Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, 58(1), 7–19.

*Putnam, H. (1975). The meaning of "meaning". In *Mind, Language and Reality: Philosophical Papers, Vol. 2*. Cambridge University Press.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.

*Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press. [Trad. it.: *La via di mezzo della conoscenza*, Feltrinelli, 1992.]

XIII. Matematica e teoria dell'informazione: testi di riferimento

Amari, S. I., & Nagaoka, H. (2000). *Methods of Information Geometry*. AMS & Oxford University Press.

*Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38(1), 173–198.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. [Liberamente disponibile all'indirizzo inference.org.uk/mackay/itila/.]

*Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.